AD-753 094

SOME TECHNIQUES FOR THE EVALUATION OF TECHNICAL TRAINING COURSES AND STUDENTS

Arthur I. Siegel, et al

Applied Psychological Services, Incorporated

Prepared for:

Air Force Human Resources Laboratory

February 1972

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151

AFHRL-TR-72-15

AIR FORCE HUMAN RESOURCES LABORATORY

SOME TECHNIQUES FOR THE EVALUATION OF TECHNICAL TRAINING COURSES AND STUDENTS

By

Arthur I. Siegel
Brian A. Bergman
Philip Federman
Applied Psychological Services, Inc.
Wayne, Pennsylvania

Wayne S. Sellman, Capt, USAF

TECHNICAL TRAINING DIVISION
Lowry Air Force Base, Colorado

February 1972

Approved for public release; distribution unlimited.

Reproduced by
NATIONAL TECHNICAL
INFORMATION SERVICE
U S Department of Commerce
Springfield VA 22151

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS

145

NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility noı any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY *(Corporate author)*

Applied Psychological Services, Inc.
Wayne, Pennsylvania 19087

2a. REPORT SECURITY CLASSIFICATION

Unclassified

2b. GROUP

3 REPORT TITLE

SOME TECHNIQUES FOR THE EVALUATION OF TECHNICAL TRAINING COURSES AND STUDENTS

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

5 AUTHOR(S) *(First name, middle initial, last name)*

Arthur I. Siegel
Brian A. Bergman
Philip Federman
Wayne S. Sellman

6 REPORT DATE

February 1972

7a. TOTAL NO. OF PAGES

137

7b. NO OF REFS

11

8a. CONTRACT OR GRANT NO

F41609-71-C-0025

b. PROJECT NO 1121

c. Task No. 112103

d. Work Unit No. 11210304

9a. ORIGINATOR'S REPORT NUMBER(S)

AFHRL-TR-72-15

9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)*

10 DISTRIBUTION STATEMENT

Approved for oublic release; distribution unlimited.

11 SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY

Air Force Human Resources Laboratory
Technical Training Division
Lowry Air Force Base, Colorado 80230

13 ABSTRACT

This handbook attempts to present methods, concepts, and considerations to be held in mind in planning and implementing a student measurement or training evaluation program. Techniques are presented, procedures are discussed, and computational examples are included. The text places principal emphasis on basic techniques, but certain more advanced approaches are also considered.

Ib

Unclassified

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
training						
education						
testing						
course evaluation						
student measurement						
quantitative methods						
training analysis						
learning assessment						
training experimentation						
educational resources						
learning experiments						
measurement						
attainm 'nt measurement						
individual differences						
training studies						
IC						

Unclassified

Security Classification

AFHRL-TR-72-15 February 1972

SOME TECHNIQUES FOR THE EVALUATION OF TECHNICAL TRAINING COURSES AND STUDENTS

By

Arthur I. Siegel
Brian A. Bergman
Philip Federman

Applied Psychological Services, Inc.
Wayne, Pennsylvania

Wayne S. Sellman, Capt, USAF

Approved for public release; distribution unlimited.

TECHNICAL TRAINING DIVISION
AIR FORCE HUMAN RESOURCES LABORATORY
AIR FORCE SYSTEMS COMMAND
Lowry Air Force Base, Colorado

Id

FOREWORD

This research represents a portion of the exploratory development program of the Technical Training Division, Air Force Human Resources Laboratory. The work was documented under Project 1121, Technical Training Development; Task 112103, Evaluating Individual Proficiency and Technical Training Programs, and was completed during the period June 1971 through March 1972. Dr. Marty R. Rockway was the Project Scientist and Captain Wayne S. Sellman was the Task Scientist. The services of Applied Psychological Services, Wayne, Pennsylvania, were obtained under Contract F41609-71-C-0025 of which Dr. Arthur I. Siegel served as principal investigator. Captain Wayne S. Sellman was the Air Force technical monitor.

Other reports prepared under this contract include.

AFHRL-TR-72-3,	Training Evaluation and Student Achievement Measurement: A Review of the Literature
AFHRL-TR-72-17,	A Survey of Student Measurement and Course Evaluation Procedures within the Air Training Command.

This report has been reviewed and is approved.

GEORGE K. PATTERSON, Colonel, USAF
Commander

ABSTRACT

This report presents a number of methods, concepts and considerations which may be used in planning and implementing a student measurement or course evaluation program for Air Force Technical Training. Techniques are presented, procedures are discussed, and computational examples are included. The text places principal emphasis on basic techniques, but certain more advanced approaches are also considered.

ERRATA

Siegel, A.I., Bergman, B.A., Federman, P., & Sellman, W.S. *Some techniques for the evaluation of technical training courses and students.* Lowry AFB, Colo.: Technical Training Division, Air Force Human Resources Laboratory, February 1972. (AFHRL-TR-72-15)

Page ii, lines 16, 17, and 18
Delete reference to "AFHRL-TR-72-27, A Survey. . .Command."

Page v, lines 21, 22, and 23
Delete sentence 2 under Results "Examples are also. . . model."

Page 8, last paragraph, line 1, should read:
1. Reliability -- a measure is reliable to the extent . . .

Page 15, line 11, should read:
. . . (armchair) rather than empirical . . .

Page 18, second paragraph, line 3, should read:
tent should be

Page 21, line 18, should read
Predictive Validity--evaluated by showing . . .

Page 65, last paragraph, line 1, should read:
. . . forced-choice item . .

Page 75, line 11, should read.

$$\sigma = \sqrt{\frac{\Sigma(x-\bar{x})^2}{N}}$$

Page 75, line 16, second and third column headings. should read:
$\underline{(x-\bar{x})}$ $\underline{(x-\bar{x})^2}$

Page 77, line 4, second and third column headings, should read:
$\underline{(x-\bar{x})}$ $\underline{(x-\bar{x})}/\sigma$

Page 92, first paragraph, last line, should read:
perimental control is .

Page 96, first paragraph, line 3, should read
utilizes the predictor-criterion . . .

Page 102, first paragraph, line 2, should read:
best fitting curve rather . . .

Ia

SUMMARY

Siegel, A.I., Bergman, B.A., Federman, P., & Sellman, W.S. Some Techniques for the Evaluation of Technical Training Courses and Students. AFHRL-TR-72-15. Lowry AFB CO: Technical Training Division, Air Force Human Resources Laboratory, February 1972.

Problem

The purpose of this report is to present methods, concepts and considerations to be held in mind in planning and implementing a student measurement or course evaluation program for Air Force Technical Training.

Approach

Selected state-of-the-art information technology is presented for use by Air Force training managers, instructors, and training evaluation specialists in developing student measurement and course evaluation programs for Air Force Technical Training. The report contains descriptions of relevant evaluation techniques, discussion of procedures, advantages and disadvantages of each technique, recommendations for application and computational examples. Principal emphasis is placed on basic techniques, but certain more advanced methods are also included.

Results

Basic techniques for student measurement and training evaluation are presented in detail along with practical applications for their use using a problem-solution format. Examples are also provided which demonstrated how these techniques can be used to evaluate each step in the Instructional System Development (ISD) model. Among the areas discussed are test construction and validity, various testing methods, performance tests and job sample tests, ratings and related methods, various quantitative methods, and more advanced statistical methods.

Conclusions

Systematic, comprehensive, state-of-the-art information on evaluation procedures is presented in an integrated useful way. Areas in the training process in need of improvement can now be more effectively identified and improved, resulting in more efficient, cost-effective training and thus more job effective graduates. The techniques presented in this paper emphasize the importance of training evaluation as a tool for collecting information to determine the extent to which training progress has occurred and training objectives achieved.

This summary was prepared by Gary G. Miller, Technical Training Division, Air Force Human Resources Laboratory.

Preceding page blank

TABLE OF CONTENTS

TABLE OF CONTENTS (cont.)

TABLE OF CONTENTS (cont.)

TABLE OF CONTENTS (cont.)

TABLE OF FIGURES

TABLE OF EXHIBITS

TABLE OF EXHIBITS (cont.)

TABLE OF EXHIBITS (cont.)

CHAPTER I

INTRODUCTION

This report describes some basic techniques for student measurement and training evaluation which appear to have particular relevance for Air Force Technical Training. The results of training evaluation and student measurement are required for: updating and improving a training program, determining if course objectives are met, determining cost-effectiveness, estimating student achievement, identifying students in need of special assistance, pointing out essential and nonessential instruction, and determining if graduates meet job performance requirements in the field.

This report is designed for use as a general source document by personnel involved in developing technical training evaluation and student measurement programs. It is not intended to be a step by step procedural "cookbook" nor is it intended to provide exhaustive coverage of the state-of-the-art in measurement and evaluation technology. However, it should prove useful as a general reference since the techniques presented are keyed to the steps in the Instructional System Development (ISD) model (AFM 50-2) and examples are provided which illustrate ways of evaluating each ISD step in order to develop improved training courses.

Systems Approach to Training Program Development

The systems approach to training is not a recent innovation. In the Air Force, the history of the systems approach, in general, dates back to the 1950s, when research and development organizations were concerned with weapon systems. A system is a complete unit of equipment and personnel (resources) who work together for a common purpose and are tied together through a communication network. AF Manual 50-2, Instructional System Development, discusses the background and history of the systems approach to training. It points out that the systems approach does not only involve curriculum planning in a logical and organized manner (systematically), but that it also focuses on objectives or ends to be achieved.

The Instructional System Development (ISD) model, as developed by the Air Force, consists of five steps:

1. Analyze System Requirements--identify the tasks to be performed within the overall environment of the operational system

2. Define Education or Training Requirements--determine the tasks that require instruction, the level of proficiency to be developed in students, and the resources needed to conduct the the instruction

3. Develop Objectives and Tests--identify behaviors required for successful job performance and construct criterion and enabling objectives, as well as achievement tests

4. Plan, Develop, and Validate Instruction--select instructional methods, media, and equipment that best satisfy learning objectives; determine the sequencing of the instructional material; validate instructional materials to prove that they teach what they are designed to teach and to insure that all elements of the instructional system function effectively in achieving stated objectives

5. Conduct and Evaluate Instruction--identify problem areas and corrective actions needed in order to satisfy the requirements of the operating commands

Advantages of Systems Approaches

The advantages to be gained by using a systems approach in training programs, development, implementation, and evaluation are increased: (1) comprehensiveness, (2) job relatedness, (3) flexibility, (4) practicality, (5) validity, and (6) appropriateness.

Disadvantages of Systems Approaches

The disadvantages of using a systems approach to training program development are its costliness and time consumption. Because of its comprehensiveness, it will naturally take more time and money to follow a systematic approach to training. The advantages and future gains, though, outweight these disadvantages.

Training Evaluation

Training evaluation is the process of making decisions and judgments about the worth or value of a training program, or parts of a training program. Secondarily, training evaluation determines whether or not training objectives have been met. The results of training evaluations serve as the basis for revising and updating training programs. In order to obtain results that will be maximally useful, studies or experiments of various types are required.

Training evaluation studies are typically conducted under two different settings: internal and external. Training evaluations performed internally are conducted within the training environment; external training evaluations are performed outside the training environment and are often referred to as field evaluations. Although training evaluations may not always provide conclusive proof of the adequacy and efficiency of an instructional program, they can identify areas for needed corrective action.

There are two concepts that are mentioned with some regularity in discussion of evaluations. These are the concepts of "high" and "low" forms of evaluation and "formative" and "summative" evaluation. In the "high" form of training evaluation, the results can be generalized across schools and situations. In the "low" form of training evaluation, though, the results are restricted to the specific research situation. Why is this so? The reason is that in the "low" form of training evaluation, the experimental conditions (which, for example, could be two different teaching methods) are not random samples drawn from the universe of all possible experimental conditions. The "low" form of training evaluation, then, is that in which the evaluator or instructor, himself, selects the experimental conditions in a restricted rather than a representative fashion. When the experimental conditions are not representative, inferences can only be made about the differences among the experimental conditions that were actually used in the experiment. Thus, if you were to examine the effects of two different methods of teaching a given subject--lecture versus motion pictures --you would have to restrict all your inferences to the subject matter content and specific motion picture and lecture types used in the experiment. However, if the experimental conditions were representative of all subject matter and teaching approaches, then generalizations across all conditions could be made. The "high" form of evaluation results when the experimental conditions are fully representative of the universe of conditions.

The other dichotomy is "formative" and "summative" evaluation. "Formative" evaluation addresses itself to the development of a training program. A "summative" evaluation is concerned with the evaluation of a program in its final form. Some people think the term "summative" evaluation is a misnomer, since "formative" evaluation never ends for instructions and training program developers. A training program is "summative" only for someone who is outside the program and looking in for a statement of its effects.

Uses and Importance of Training Evaluation

Training evaluation can be used for acquiring data relative to course planning and administration, including student classification, diagnosis of learning disabilities, appraisal of student progress, and student advancement. Training evaluation can also be used for assessing the effectiveness of instruction.

Included within these uses are: assessing changes in student behavior, determining whether training achieves its stated objectives and goals, evaluating techniques and personnel, detecting and correcting behavior problems, modifying teaching procedures when considered appropriate, determining whether desired achievement levels have been reached, and determining the progress, course, and extent of learning.

In the ultimate sense, training evaluation is used to determine whether to modify, keep, or end a program. All of the above statements about the uses and importance of evaluation are buried within one or another context. When removed from the restrictions of context, we can say that evaluation is information which is acquired for use in decision making processes.

Student Measurement

Training programs are designed to graduate students who have satisfactorily met certain stated objectives. Through a program of student measurement, the determinations are made of whether or not a student has achieved these objectives or requirements. Since the training objectives are concerned with behaviors, student measurement, then, is the process of determining if a student has acquired the necessary behaviors. Sometimes, the extent to which the student has achieved the objectives is referred to the relative standing of the student within a group of his peers. This is termed "norm referencing." Sometimes student achievement is based on some absolute standard. This is termed "absolute" or "criterion referencing." In the Instructional Systems Development Course: Abbreviations and Glossary of Terms manual (ATC handout 3AIR 75130-X-1, 15 September 1970), measurement is defined as: "The process of determining as objectively as possible a student's achievement in relation to other students or to some fixed standard in order to determine a formal grade" (p. 10). Student measurement can also involve, as an artifact, the evaluation of the strengths and weaknesses of an individual for diagnostic and subsequent remedial action and generation of information relative to inferences about the effectiveness of the instruction. This last statement defines an area in which student measurement and training evaluation overlap.

Uses and Importance of Student Achievement Measurement

Some of the various uses of student measurement are:

1. feedback, diagnosis, and steering of the student
2. helping the student to plan and evaluate his own educational experiences
3. establishing merit
4. determining if learning has occurred
5. evaluating the effectiveness of training
6. determining whether a student can perform adequately on the job, or in a higher level training program.

All of these uses, more or less, mirror the definition of student measurement. The definition of student measurement, then, specifies its use. Also, to some extent, the definition defines its importance.

In summary, student measurement is important for determining the adequacy and current skill levels of trainees within a training program.

CHAPTER II

DIMENSIONS OF TRAINING EVALUATION

The systems approach to training program development, as defined by the ISD model, attempts to account for all variables that can affect training and student behavior. Courses developed on the basis of this model begin with a job analysis for the purpose of determining behaviorally oriented requirements. Then, decisions are made regarding the behaviors which require training in the course (certain skills may be reserved for on-the-job training because they are easy to learn or are very rarely called for when performing on the job). Next, training objectives are specified and tests developed; the course is then planned and the instructional methods, media, and equipment are determined and the course is conducted. Finally, an evaluation of the course is performed to determine the extentto which course graduates meet training objectives, the effectiveness of the instructional methods, training literature, etc. Additionally, an evaluation of the nonspecified outcomes of the instruction may be conducted. Such an evaluation considers by-products of the course and could lead, for example, to indications that graduates were positively affected by the instruction if they were stimulated to extra reading on the subject matter. On the other hand, an undesirable outcome of instruction would be indicated if graduates profess a dislike of the subject matter. Undesirable intangible outcomes of instruction indicate problem areas in the training program and needed areas for revision.

Concepts and Considerations

Evaluation Attributes

A training evaluation, to be complete, should include consideration of the students' knowledges, skills, and attributes as they relate to both the training objectives and the performance requirements. Also, the evaluation should determine whether or not the elements of the training program are current, up to date, and effective and efficient from the point of view of meeting the educational or training requirements. Training evaluation studies should also pay very close attention to over and undertraining relative to each objective. Overtraining represents a negative feature in a cost effectiveness evaluation, while undertraining indicates less than adequately prepared graduates.

In order to have a robust training evaluation study, the study must not involve those weaknesses which have appeared in so many evaluations. To avoid these problems, guidelines must be drawn up in advance. They must describe fully the methods, procedures, data collection instruments, and statistical methods to be employed. Estimates of time needed for conducting each aspect of the study should also be included. The guidelines should be as specific as possible and not subject to personal interpretation.

Aside from the development of concise and specific guidelines, evaluations should: not overstress undocumented subjective opinion, standardize performance measurements, exercise control in terms of personnel and equipment changes, and describe adequate and precise criteria. Moreover, since a set of recommendations from an evaluation can be no better than the data on which the recommendations are based, a number of data considerations must be held in mind.

Training Evaluation Data Considerations

Decisions and judgment have to be made on the basis of training evaluative data. For this reason, the data must be weighted against a variety of criteria. In order for a data set to be scientifically defensible, the measures on which the data set is based should meet most, if not all, of the following criteria:

1. Reliability--a measure is reliable to the extent that it will yield a similar score when a person is measured with it on two separate occasions. If a measure is not capable of doing this, it is worthless for evaluation purposes. There are several methods for assessing the reliability of an evaluative measure and all of them involve the calculation of a correlation coefficient. Any elementary statistics textbook will instruct the reader in methods for calculating and interpreting a correlation coefficient. Chapter III of this

manual also discusses and describes some of the simpler methods for calculating a correlation coefficient.

2. Validity--Validity refers to the extent that a measure assesses what it intends or purports to measure. In the particular context in which we are speaking, the measure must have most of its content in common with the training objectives. If the measure does not have content in common with the training objectives, then it is evaluating someing other than what is intended. An evaluation instrument can be highly reliable, and yet have little or no validity.

3. Comprehensiveness--Any useful evaluative measure must sample from the total range of training objectives and not from a portion of them. Otherwise, data pertinent to some required decisions will not be on hand and necessary tradeoffs will not be possible.

4. Objectivity--Only those aspects of student behavior whose measurement is not dependent on the judgment or accuracy of the observer should be included in the measurement instruments. If the evaluator's personal biases can enter into the data, he is being asked to be subjective rather than objective. A measure is objective to the extent that different evaluators who use the same measure will give a student the same score.

5. Differentiation--Evaluation measures must be able to reflect differences in the variable being measured. This allows decisions and comparisons to be made on the basis of known amounts by which one group differs from another. The difference need not be a score in the usual sense; it may be the frequency with which certain behaviors are found or even just the number of persons in each of two groups.

6. Relevance and Appropriateness--One must select measures which are most appropriate for the objectives in question. A simple method one could use for this purpose is to construct a matrix or grid. If you were evaluating the students' mastery of the subject matter, then down the left side of a sheet of paper you would list the training objectives. Across the top of the page, you would list the various ways through which achievement of the training objectives could be measured. Some of these might include: oral tests, written tests, and performance tests. One of these, or some other type of measure is likely to be most appropriate. For each objective you simply check off the appropriate measure. Costs are another consideration. They are discussed under item 8 below.

7. Correct Weighting of Elements--If a set of measures is employed and a total score is to be derived across these measures, then each of the subscores must be weighted in terms of its relative importance. Moreover, if scores are to be combined they must be based on a common metric (e.g., based on standardized scores). Methods for standardizing scores are discussed in Chapter IV of this report.

8. Cost/Effectiveness--The evaluator must select the measures which will best meet the prior criteria for the least cost. One does not select a very sophisticated method or measure when a simpler one will do just as well. Many times cost/effectiveness determination reduces itself to a question of fidelity. You may wish to measure the trainee's performance on the actual piece of equipment he will be using on the job. Many times, though, this is too costly. A measure can possess complete fidelity (e.g., be based on behavior measured on an actual piece of working equipment or in an actual job situation), or possess partial fidelity (e.g., be based

on a written description or drawing of the equipment or situation). Some of these latter low fidelity measures, although inexpensive, may give information which is equally adequate with a sophisticated simulation for the purposes on hand. Often, the level of sophistication of the measure used will vary with the job or training in question.

Data Considerations

When performing an evaluation study, the researcher should plan ahead for the required data analysis. One should not go out and perform an evaluation study with 1,000 graduates if a computer and a computer programmer are not available to assist in the data analysis. In other words, do not plan to collect more data than you can analyze.

Another consideration is the number of subjects you will use in your evaluation study. Many times the merit of your conclusions will be directly dependent on the number of subjects involved. A correlation coefficient of .50 will not be statistically significant beyond chance for 10 subjects, but it will be significant beyond one chance in 100* when it is based on 40 subjects.

Training Criteria

A criterion is a standard or basis for making a judgment, e.g., a test score of 63 on a certain test is a standard for seventh graders as based on the criterion of the performance of the seventh grades in 1,000 public schools. Training criteria can be divided into two groups, internal and external. Internal criteria are directly concerned with the training itself, while external criteria measure posttraining or on-the-job behavior.

*A correlation coefficient is considered to be statistically significant, by convention, if it can occur by chance five times out of 100. Statistical confidence levels are usually denoted as .05 for five chances in 100 and .01 for one chance in 100.

Use of an overall or composite criterion will almost always conceal important relationships since many of the subcriteria within the composite are independent from each other. It is preferable, then, to use multiple criteria which reflect different aspects of the behavior being studied in order that required behavioral data will not be submerged in an overall score.

The criteria must not be affected by the method of measurement or research procedure. Even the presence of the experimenter or the process of evaluation itself can alter the results.

One of the more important criteria for evaluating training programs is the amount of transfer of learning from training to the real job situation. When transfer of training criteria are not available, then intermediate criteria can be used (e.g., a final course examination).

In training evaluation and student achievement measurement, the testing of terminology (which is specific to the training course) should be kept independent from tests of the understanding of course content. A person who is not taking the course should be able to understand (not necessarily answer) every question on the knowledge test.

Generally, there are two types of training measures: interim and terminal. These training measures are different from selection measures which are administered prior to instruction or training. Selection measures are used for establishing initial levels and for classification purposes. The correlation between selection tests and future performance should be high. The training measures include:

1. Interim Measures--measures taken while training is in progress. These are usually better predictors of final performance than initial measures.

2. Terminal Measures--measures obtained after training is completed. These are predicted by the initial and interim measures. Some examples of terminal measures are: written tests, oral tests, performance tests, instructor evaluations, and rating scales.

Evaluation Problems

Too often evaluation studies focus more on the measurement of trainee reactions to the exclusion of trainee learning, trainee behavior on-the-job, and the effects of the training on the organization. Also, much innovation in training is done for its own sake (e.g., the relief of boredom) and only secondarily for its outcomes. Additionally, evaluation studies are too often large scale and aimed at funding agencies to prove that the innovation is of value.

Evaluation can have both positive and negative effects. The student being evaluated will always respond to the evaluation in terms of its perceived fairness. If he perceives the evaluation as unfair, the student may become resentful, especially if the evaluation is threatening to his career or to his status.

Moreover, it is difficult to complete an evaluative study in an environment which resists social change. If the administration perceives the evaluation as negatively affecting vested interests, many obstacles can be placed in the evaluator's path. Each of these, in itself, may be minor but the integral may be such as to prevent study completion or to weaken the emergent conclusions to the point that they are meaningless.

When reporting, evaluation studies, one must consider the type of person or organization that will review the findings. These different types are trainers, curriculum planners, administrators, sponsoring organizations, learning theorists, etc. The value of a report to a particular person will depend on his needs, and the report must be tailored for the specific user. Sometimes, different reports may be required for each user.

Exhibit 1 presents a checklist which can be employed to help to avoid a number of other problems which have sometimes been associated with training evaluation study planning and conduct in the past.

Exhibit 1

Training Evaluation Plan Checklist

1. Have I selected the problem arbitrarily or in view of user needs or on some other logical basis?

2. Does my study plan overstress resources and materials and understress performance effectiveness?

3. Does the plan stress quantity of services and record keeping at the expense of true evaluation?

4. Does the plan emphasize program objectives which are based on tradition rather than on systematic development through the ISD model?

5. Have I avoided mixing of final, intermediate, and immediate objectives?

6. Is idealism emphasized at the expense of realism?

7. Have I emphasized collection of data from available or existing records and avoided collection of new data?

8. Is the plan based on a study design which will allow me to know whether the result is due to chance?

9. Are my measurement methods reliable? accurate? objective?

10. Are my criteria comprehensive? relevant? correctly weighted?

11. Are my weighting methods based on empirical rather than rational methods?

12. Have I planned for any required training of the persons who will conduct the evaluation and for standardized conditions where required?

13. Does the plan make allowance for or consider controlling demographic (e.g., race, age, education) and locational variables?

Exhibit 1 (con.)

14. Does the plan overemphasize the collection of self evaluational data which are easily biased?

15. Does my plan involve the use of supervisors to collect the required data? If so, have I planned arrangements to relieve them of some of their regular duties so that they may give sufficient attention to the evaluation?

16. Do equally powerful but more cost/effective methods exist for achieving the same result?

We can sum up by saying that training evaluation has been characterized by too much use of rational (armchair) rather tham empirical evaluation methods. Similarly, evaluation research has been frequently subjective when objectivity was needed. Finally, evaluation research has too often been limited by monetary considerations. The monetary criticism is probably the most important, since many of the other criticisms can be reduced to it. What investigators often fail to realize is that cost cutting actually wastes money because the results of inadequate evaluations are, at best, uninterpretable and, at worst, misleading and invalid. Many agencies, contractors, and others performing evaluation studies might be well advised to save their money or to perform one or two sound evaluation studies rather than the five or six poor ones.

Lastly, some investigators claim that broad based evaluation programs have design and technical problems so ponderous as to make any evaluation impractical and questionable. In situations where there are many variables to consider, one cannot possibly prove or disprove the value of any program.

CHAPTER III

DEPENDENT MEASURES

In research, a response is always considered the dependent variable, and it is the response that is observed and measured. (The independent variable, in research, is the variable that is manipulated to see what effect changes in that variable will have on the dependent variables.) In a training program, the dependent variable could be the trainees' responses on an achievement test. (An example of an independent variable in such a setting might be the training method.) Training evaluation criteria can be divided into those based on the content of the training program (internal criteria), and those based on job behavior (external criteria).

There are two kinds of conditions which can indicate that learning has occurred. In the fixed condition, a response or instance of behavior is used to show that learning has taken place. The student's answer is the criterion of performance. In the variable condition, several responses can show that learning has occurred. One can easily see that, within this latter condition, it is easier for the student to demonstrate that he understands the concept being taught, since the requirement for learning in the variable condition is dependent on a concept or rule and not on a response. This does not mean we should always use the variable condition. The fixed condition is, at times, also needed. The variable condition is more often experienced in performance tests or criterion referenced lists where many actions and responses are performed. The simple response or behavior (fixed condition) is more often tested in a written test where a single answer is sought for a specific point.

Exhibit 2

Examples of Independent and Dependent Variables

Independent Variables	Dependent Variables
Class size	Final examination
Student mix	Gain
Training time	Rate of learning
Instructor style	Class standing
Social milieu	On-the-job performance
Student personality	Pass or fail course
Prerequisites	Instructor's report

Preceding page blank

Test Construction

The important consideration in constructing tests is that the tests measure the achievement of specific objectives. The objectives are usually stated in behavioral terms and as such can be readily measured.

In planning the test (but following the identification of course objectives, the lessons, and the testing objectives), the course contents should be outlined. This outline need only contain the major categories or headings. So that the test will measure a representative sample of learning objectives and course content, a table of specifications is recommended. A sample of such a table is presented in Exhibit 3. This example represents a portion of a specification table for a course for an introductory block of the Air Force Materiel Facilities Specialist course.

The numbers in each cell of the table indicate the number of test items to be written in that area. As such, the numbers in each cell represent a weighting scheme. But, the weighting scheme should be decided on before constructing this table so that the number of items to be given to each area is known in advance. The considerations that go into the selection of the number of items to include in the various areas are: the importance of each area in the total learning picture, the amount of time devoted to each area in the course, and the objectives that have the greatest value.

The test items are then written. Sufficient items are written so that the requirements of the specifications table are surpassed. An item pool of more than the finally required number of well written items should be developed so that if replacement items are needed for the test, they will be available.

There are many different types or forms of test items that can be used. These are discussed in a later section of this chapter. Before selecting a particular item form, the advantages and disadvantages of each item form for the purposes on hand should be considered. Additionally. the nature of the learning objective to which the item is aimed should be considered. A test item should measure the learning objective as directly as possible and, depending on purpose, certain item types are better than others. Multiple-choice items appear to be superior to other types of written test items. They can measure a variety of learning objectives from simple to complex.

Exhibit 3

Examples of Specifications for Test of Block I, Materiel Facilities Specialist Course

	Content			
Objectives	Principles	Procedures	Use of Equipment	Total
Can describe purpose and structure of AF supply system	10	0	0	10
Can operate UNIVAC 1050-II computer	5	15	5	25
Can use and update Supply Manual, AFM 67-1	10	20	0	30
Can arrange storage facilities according to standard principles	10	0	10	20
Can enter information on condition tags	0	10	0	10
Etc.	etc.	etc.	etc.	etc.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
TOTAL	25	50	25	100

The multiple-choice item, along with true-false and matching items, are referred to as "selection" test items because the student selects his response from alternatives that are offered to him. The selection test items have certain advantages over "supply" items, i. e., those items for which the student has to furnish his own response as in the essay item. The advantages of the selection type item are that: (1) test scorer personal biases and subjective idiosyncracies are removed from the scoring, (2) ease of scoring (scoring keys or machines can be used), (3) more aspects of a course can be tested in a given period of time, (4) with increases in comprehensiveness improved validity may be expected, and (5) ease of application of statistical item analysis procedures to resultant data.

After writing the items, they are ready to be assembled as a preliminary test. There are certain rules to follow in arranging test items. These are: (1) items of the same type should be grouped together, (2) items should be arranged in order of increasing difficulty, and (3) items should be grouped together which measure the same learning objective or subject matter content. When each item is numbered in the test, its number can be placed in the appropriate place in the specification table so that a record of the test area contents is maintained.

The directions for the test should be simple and concise, yet complete. Instructions on the manner in which the student is to respond and the tools and materials he can use in the test should be provided. The procedure to be used in scoring the test should be explained to the students. Score sheets and score keys should be prepared so that the actual scoring of the test will be facilitated.

The test is then administered to a sample which is representative of those on whom the test will eventually be used and scored. An evaluation of the effectiveness of each item, an item analysis, is next performed. The item analysis provides information on the difficulty of the item (proportion of students who answer the item correctly), the effectiveness of each alternative (proportion of students who answer the item incorrectly and who are attracted to each distractor), and the item validity (do the students who score high on the total test answer the item correctly?).

Items which do not meet appropriate criteria, as indicated by the item analysis, are rewritten or new items are substituted for these. Of course, when an item is substituted or altered, the substituted or altered item must be subjected to the same item analytic procedures. General criteria for acceptable items are:

difficulty level	=	.40 to .60
item validity	=	correlation coefficient greater than .20
distractor effectiveness	=	equal spread of wrong answers over all distractors within a multiple-choice question

Test Validity

The validity of a test or the extent to which it measures whatit is supposed to measure can be assessed in several ways. There is no such thing as general validity. Validity is always validity for a particular purpose.

There are four kinds of validity:

Validity--evaluated by showing how well a test predicts subsequent behavior. End of course grades in a mechanics course, say, should indicate a student's competence in maintaining and repairing vehicles. If students with high grades perform better on-the-job, we would say the end of course grades are predictive of on-the-job success, or that they have predictive validity. We could objectively measure this validity by computing the correlation between grades and, say, supervisor's ratings of job performance. We would call such a correlation a validity coefficient.

Construct Validity--evaluated by showing that a test actually measures the trait or ability it was intended to measure. Essentially, in construct validity, it is the theory underlying the test that is being validated and is one of the most difficult kinds of validity to confirm in practice. Three of the ways it might be done are described as follows:

Experimental manipulation. Here we would try to experimentally increase or decrease the strength of the trait. An example would be showing that scores on a test of anxiety increase under stressful circumstances.

Measuring naturally occurring groups. If there were groups of people that could be expected to vary naturally on the trait, test scores should vary in a similar manner. An example would be to administer a test of managerial aptitude to a group of company presidents, supervisors and clerks.

Correlation with similar tests. A frequent devise to demonstrate construct validity is to show that the test correlates significantly with a test generally recognized to measure the construct in question.

Concurrent Validity--evaluated by showing that scores are similar to scores from other testing procedures. An example would be computing a correlation between paper and pencil derived scores and scores derived from actual "hands-on" performance. If the correlation was high the test would be said to have concurrent validity in the sense that it was highly correlated with similar measures. In this case the paper and pencil test might be considerably cheaper to administer.

Content Validity--evaluated by showing how well the test items actually sample the range of behavior necessary to show subject mastery or competence. There is no way to actually measure content validity except by showing that items which constitute the test in fact cover the subject content. This can be achieved by performing a detailed analysis of the subject objectives and subobjectives and then constructing items which measure mastery of these objectives. Content validity is related to the notion of face validity.

Face Validity--evaluated by simply inspecting the items for their obvious relevance to the subject matter. An item asking the examinee to spell a word would lack face validity for a mathematics test. Even though items lacking face validity may be shown to have good predictive validity (and therefore value), they may cause examinees to question the seriousness of the test and might therefore produce less cooperation.

Hierarchical and Sequential Testing

Hierarchical and sequential tests involve a branching sequence in which the student only receives items at his own level. If a student answers a test item incorrectly, his next item will be an easier one, but if he answers correctly the next item will be more difficult. The concept of sequential testing was introduced early in intelligence tests, and in recent years has been used in achievement tests, as well. The procedure has the following advantages: (1) it decreases test time (especially for students at the extremes of the distribution because they can be routed quickly), (2) it increases reliability, and (3) it increases student motivation because the student is not forced to take and guess at more difficult items.

There are both one stage and two stage procedures in sequential testing. Two stage procedures have a routing section which branches the student to the appropriate items and a measurement section containing items of suitable difficulty. One stage procedures use a routing stage only. These procedures both use fewer items for those persons easy to classify and more items for those persons at the borderline of categories. Computer based testing facilitates this procedure.

A specific example of the sophisticated methodology used in developing and administering a sequential testing procedure is beyond the scope of this manual. Most of the methods now in operation require consultation with test experts and the use of programming. A simplified account of the approach will be presented, though, so that the reader will be able to determine if sequential testing can be applied to his own evaluation and measurement situation.

The specific steps in the method are:

Test Development

1. collection of several hundred items representative of the subject matter area in question.

2. administration of the items to a pool of 50-100 subjects.

3. determination of the proportion of subjects in the sample who answer each item correctly (from 0.00 to 1.00). This is the item difficulty level.

4. selection of groups of items representing the .05 -.15, .25 -.35, .45 -.55, .65-.75, and .85 -.95 difficulty levels. The average difficulty level in each of these item groups should be .10, .30, .50, .70, and .90.

Test Employment

5, administration of the .50 level items to all the students being evaluated. Students who obtain 40 to 60 per cent of these items correct are at the .50 hierarchical level.

6. administration of the .90 level item group to those subjects who correctly answer at least 90 per cent of the .50 level items.

7. all students who correctly answer 70 to 80 per cent of the .50 level items take the .70 level items.

8. students correctly answering two or three of the .50 level items correctly then take the .30 level items.

9. finally, students who answered none or one of the .50 level items correctly take the .10 level items.

Whenever 40 to 60 per cent of the items in an item group are answered correctly, the student's hierarchical level has been determined. Students answering fewer than 40 per cent of the items correctly are routed to items at the next lower level, unless they are already at the lowest level. Students answering at least 70 per cent of the items correctly are routed to the next higher level, unless they are already at the .90 level.

A check that the sequential test was adequately constructed can be made by comparing the scores of students on a higher level item group with their scores on the next lower item group. For example, if a student answers more items correctly on the .70 level than on the .50 level, then the levels are not truly hierarchical and the test must be modified.

This method of sequential testing can be used with single items representing different hierarchical levels, as well as with item groupings. When a student answers an item correctly in the single item method, he is routed to an item at a higher difficulty level. Conversely, if a student answers an item incorrectly, he is routed to an item at a lower hierarchical level. In the present sense, the student is continually routed until he reaches a level above which he answers none of the questions correctly and below which he answers all of the questions correctly. However, in the more practical circumstance of today's classroom, certain allowances have to be made because it is not easily possible to develop an achievement test such that there will be a group of questions that nobody can answer correctly.

In summation, the advantages of sequential testing are:

1. the student receives test items at an appropriate difficulty level
2. the student is not frustrated by difficult items
3. test taking time is reduced
4. reliability is increased
5. student motivation is increased
6. fewer test items are needed
7. computer based testing can be used

One disadvantage of the sequential testing procedure is that the development of such programs requires sophistication and personnel outside the scope or availability of most instructors and student measurement specialists. Moreover, the method often works best for persons at the extremes of the ability range. The closer the examinee is to the actual mean, the more items will be needed in his test to arrive at his absolute upper limit. This is actually a limitation rather than a disadvantage.

Criterion and Norm Referenced Testing

When a student takes a criterion referenced test, his results are compared with an absolute standard to determine whether or not he has attained acceptable (criterion) performance. The absolute standard may be passing a performance item which is at a given level of difficulty or achieving an absolute score. By contrast, the students' results on a norm referenced test are compared with the performance of a peer or reference group, e.g., a given percentile.

The characteristics of criterion referenced tests are that they:

1. indicate the degree of competence attained by an individual independent of the performance of others

2. measure student performance with regard to specified absolute standards (criteria) of performance

3. minimize individual differences

4. consider variability irrelevant

Generally, from the above, it can be seen that criterion referenced tests tell us how the student is performing with regard to a specified standard of behavior. Criterion objectives, expressed in terms of specific behaviors, are identified. They specify the operations and knowledges a student must demonstrate to satisfy a job performance requirement. Individual differences among students are considered irrelevant, since the student is graded against a single external standard rather than against all the others taking the test. It would not make sense to assign grades to students on the basis of relative performance, if it is not known whether any of the students actually attained a specified behavioral objective. Hence, in criterion referenced measurement programs, grades are usually S(satisfactory) or U(unsatisfactory).

One can, though, derive information about individual differences from criterion referenced tests by specifying the degree of competence reached by each student.

In one sense, there are no real differences between criterion and norm referenced tests; the difference lies in the method by which passing or failing scores are set.

Since an important concept underlying the systems approach to training is that of training students to a minimum standard as represented by the training objectives, it appears that the criterion referenced testing program is the most appropriate for the Air Force.

The question sometimes arises as to when it is more desirable to use a norm referenced test as compared with a criterion referenced test. The crucial consideration in this regard is the purpose of obtaining scores. If scores on an achievement test are to be used to evaluate students, one against the others, to demonstrate how well one compares with his peers, then the norm referenced test is used, but the test must contain items of different difficulty levels. On the other hand, if an indication of absolute proficiency or absolute gain is wanted, then criterion referenced tests are employed. Criterion referencing, for example, would best be used in a performance evaluation which was aimed at determining the readiness of airmen to perform corrective maintenance on a given aircraft. Criterion referencing is also frequently used when a direct relevance to the job is sought, when content validity is to be maximized, and when global (pass-fail) discrimination is sought.

Several sequential steps are involved in the construction of a criterion referenced test. An example of how a criterion referenced performance test was constructed for machinist trainees is given below:

1. after a job analysis, the dimensions of the machinist job were defined.

2. valve assembly proficiency was considered to be one job aspect necessary for an entry level machinist to be considered to meet the minimum requirements.

3. a checklist type performance test was developed to measure student ability to follow the procedures and steps involved in the assembly of a gate valve. This checklist is presented in Exhibit 4.

Exhibit 4

Valve Assembly Scoring Checklist

______________________________ ____________________
Name Date

1. Takes packing nut and stem ________

2. Screws packing nut to top of stem ________

3. Takes gate ________

4. Screws gate on bottom of stem ________
   (Prompt if Step 4 performed incorrectly)

5. Winds gate all the way up stem ________
   (Prompt if Step 5 performed incorrectly)

6. Inserts gate and stem assembly into body of valve ________

7. Screws gate and stem assembly onto body of valve ________

8. Inserts handle in top of stem ________

9. Screws handle to top of stem with handle nut ________

10. Screws on first 3/4" nipple ________

11. Screws on second 3/4" nipple ________

12. Checks assembled valve to see if parts are fitted tightly ________

Total ________

4. **after developing standardized examinee and examiner instructions, the test was administered to trainees.**

5. **the trainee was given two points for each item performed in the proper sequence. One point was given for each item performed correctly, but out of sequence. Trainees who finished the assembly in less than one minute, without error, received three bonus points.** Trainees who correctly finished the assembly in 60-90 seconds received a one point bonus. The total possible score, then, on the assembly test was 27 points.

6. job experts were consulted in order to determine the minimally adequate score for this test. Minimally adequate was defined in terms of the job requirements.

7. twenty-two points were considered to be a minimally adequate criteria referenced score.

8. hence, all trainees whose scores were 22 or better exceeded the criterion referenced score for the Valve Assembly test

9. any individual who received a score of less than 22 points was considered to fail the test.

As the reader can readily see, the criterion referencing may be adapted for paper and pencil testing. In fact, any test can be made a criterion referenced test by assigning a score which represents a specific behavioral objective.

In a field survey performed for the Air Force Human Resources Laboratory, training specialists were questioned on the advantages and disadvantages of criterion referenced tests. They responded to these queries from their experience with this type of test, in the Air Force technical training context. The general consensus was that the criterion referenced testing represents a thorough and practical method for determining if students meet training objectives. Additionally, criterion referenced testing was said to be an appropriate method for keeping track of student progress. On the other hand, the training specialists found certain difficulties with criterion referenced tests. They found them somewhat difficult to construct, costly in terms of the time to construct and administer, subjective in the scoring procedures, and to involve administrative problems insofar as the requisite equipment was concerned and the amount of time that was allotted to testing.

Mastery Testing

A variant of criterion referenced testing is known as "mastery testing." In mastery testing, the trainee must reach a certain level of proficiency or mastery in order to achieve a passing grade. In this sense, mastery testing is equivalent to criterion referenced testing.

Confidence Testing

Confidence testing involves a method which provides for weighting the selected alternative(s) of a test item, so as to allow the examinee to reflect his belief in the correctness of his response. The basic concept behind confidence testing is that there is additional information available from the students' degree of belief (confidence) probabilities. The method, accordingly, allows the student to maximize his expected score if he truly reflects the degree of his belief, or the probability that a specific alternative is correct.

Confidence testing evolved because of the feeling among student measurement experts that knowledge, as measured by achievement tests, has more dimensions than those that are indicated by the typical multiple-choice and true-false test. Some students can respond to a multiple-choice test item with 100 per cent certainty of the correct choice. Other students may be able to eliminate several alterntives as being incorrect and have to make a decision between one or the other of the two remaining alternatives. Still other students may approach the same test item and not be able to eliminate any of the alternatives. The question then arises, does the student who can eliminate all the incorrect alternatives have more knowledge than the one who could eliminate all but two? And, by the same token, does the student who is uncertain of his response (and so indicates through his statement of his confidence in his answer) have more knowledge than the one who selects the same incorrect response, but feels certain that his response was correct? Confidence testing advocates would respond in the affirmative to these questions.

The "Pick-One" method will be used as an example of the confidence testing method. A more complete account of this method can be found in AFHRL-TR-71-32 and AFHRL-TR-71-33. In the Pick-One method, the student first picks the answer he thinks is correct from among the other multiple-choice alternatives. The student then assignes a probability value indicating his confidence in that answer. This probability value is then converted to an item score by using Exhibit 5. For example, assume that a student assigns a probability of .7 to his answer. If there were four alternatives, he would receive a score of .84, if his answer was correct, and a score of -.76, if his answer was incorrect. If there were five alternatives in the item, the student would receive a score of .86 if his answer was correct, and a score of -.70 if his score was incorrect.

In summation, then, the main advantages of confidence testing are the more thorough assessment of student knowledge, and the virtual elimination of chance as a factor in test scores.

The disadvantages of confidence testing are the difficulty and time required to score the test. However, these disadvantages are probably outweighed by the advantages. Moreover, computer scoring can be employed to ease the scoring burden.

Exhibit 5

Scoring Table for Pick-One Confidence Testing*

Probability Corresponding to Selected Alternative	Alternative 2		Alternative 3		Alternative 4		Alternative 5	
	If Correct	If Wrong	If Correct	If Wrong	If Correct	If Wrong	If Correct	If Wrong
.2							0	0
.25					0	0	.12	–.04
.3					.03	–.05	.23	–.08
.333			0	0	.21	–.09	.31	–.11
.35			.04	–.01	.25	–.11	.34	–.13
.4			.18	–.11	.36	–.17	.44	–.19
.45			.32	–.21	.46	–.25	.53	–.25
.5	0	0	.44	–.31	.56	–.33	.61	–.33
.55	.19	–.21	.54	–.43	.64	–.43	.68	–.41
.6	.36	–.44	.64	–.56	.72	–.53	.75	–.50
.65	.51	–.69	.72	–.70	.78	–.64	.81	–.60
.7	.64	–.96	.80	–.85	.84	–.76	.86	–.70
.75	.75	–1.25	.86	–1.02	.89	–.89	.90	–.82
.8	.84	–1.56	.91	–1.19	.93	–1.03	.94	–.94
.85	.91	–1.89	.95	–1.38	.96	–1.17	.96	–1.07
.9	.96	–2.24	.98	–1.57	.98	–1.33	.98	–1.20
.95	.99	–2.61	.99	–1.78	1.00	–1.49	1.00	–1.35
1.00	1.00	–3.00	1.00	–2.00	1.00	–1.67	1.00	–1.50

*Adapted from AFHRL-TR-71-32

Partial Knowledge Testing

Traditionally, in scoring a four choice multiple-choice item, the student is awarded one point for the correct answer and no points for a choice of any incorrect answer or distractor. Partial knowledge exists when the student can identify one or more of the distractors. Using this technique, in a multiple-choice format, one point is given for each distractor identified and three points are subtracted if the correct answer is identified as a distractor. Scores on each four choice item can range from plus three to minus three. Partial knowledge testing, then, yields increased item and test variance and penalizes for random guessing. Two possible disadvantages to this method are that it is not applicable to all kinds of tests (e.g., true-false), and the scoring is time consuming. In addition, a personality factor enters when a person identifies two distractors and has to make a choice between standing pat on two points or risking a minus one to get three points.

Unobtrusive Measures

Unobtrusive measures are those which do not interfere with on-the-job behavior, or behavior in training. Most of the methods we have described thus far intrude on behavior by requiring the trainee to take a test or perform a task. Some of the less obtrusive measurement methods are:

1. Observation

   In this method, the supervisor or trainer simply observes the behavior of the trainee either in training or on the job. Usually some kind of checklist is used by the evaluator. A checklist, in this instance, is a list of the behaviors that are required of the trainee or job incumbent. The evaluator checks all those items on the checklist which the trainee performs correctly. The final score is the number of items checked. (An example of a checklist for an assembly task was presented on page 27.)

2. Rating and Ranking

Rating and ranking, discussed in later portions of this manual, are variations of unobtrusive measurement.

3. Film, Closed Circuit T.V., and One-Way Mirror Viewing

These are all variants of the observation method and usually employ some rating, ranking, or checklist procedure. These methods are different from direct observation in that the evaluator's presence is not known to the student. This is an advantage, because the presence of a supervisor or trainer can often cause a modification in the behavior of the person being observed.

Performance Tests and Job Sample Tests

Checklists

Much of the previous discussion was concerned with training evaluation and student achievement measurement through various applications of written tests. In a recent study, performed for the Air Force Human Resources Laboratory, student measurement specialists were asked for their opinions of this type of test. They indicated that, although paper and pencil tests might motivate students to study, such tests are ineffective for measuring achievement in the Air Force technical training courses where skills are the essence rather than knowledges. In performance tests, the person being evaluated performs tasks which are relevant to his present or future job. Some performance tests are less obviously related to jobs than others. Performance tests can range from simulated performance through performance of job tasks using actual job equipment. Scoring can be based on measurement of performance in process, adherence of a final product to prescribed standards, care and use of tools during performance, adherence to safety precautions, or some combination of these categories. Completion of a prescribed task within an allotted time is also sometimes scored. One of the most popular methods of scoring is through a sequential checklist. To construct such a checklist, a task is broken down into the sequential elements which must be correctly performed if the task is to be completed. Each of these elements is then sequentially listed for scoring by an examiner while he observes the performance of the person being tested on the task. A sample of such a checklist for "Cutting a Full Face Gasket" is presented in Exhibit 6.

Exhibit 6

Example of Scoring Checklist for a Sequential Performance Test

______________________________ ______________
Name Date

1. Places sheet gasket material over face of flange. ________
2. Uses round end of ball pean hammer, ________
3. and taps out bolt hole ________
4. lightly. ________
5. Inserts bolt in hole. ________
6. Taps out diagonally opposite bolt hole and inserts bolt. ________
7. Taps out remaining bolt holes. ________
8. Taps out inside circumference of flange ________
9. using round end of ball pean hammer. ________
10. Taps out outside circumference of flange ________
11. using flat end of ball pean hammer. ________
12. Carefully removes pieces of excess material from gasket and flange. ________

Total ________

For the test shown in Exhibit 6, two points were allowed for each step--one point for correctly performing the step and the second point for the correct sequence. The test administrator simply placed 1 or 2 in the space next to each step. Final score was determined by summing these points.

In many instances, you will be interested in the student's use of tools, his adherence to safety precautions, and the quality of his final product, as well as the manner in which he performs. A sample test employing this combined scoring for a welding performance test is presented in Exhibit 7.

Examinee Recorded Performance Tests

A variation generally included under the performance test rubric is the examinee recorded approach. This test type is useful when there is little testing time, or when many people must be tested at once. The kinds of information recorded by the examinee may include measurements, locations, and interpretations. One must be careful not to test reading and writing skills when using tne examinee recorded approach. The written response required by the examinee should be as simple as possible (e.g., checkmark, underlining, one word answer). This is especially true when administering performance tests to low aptitude personnel. Some examples of examinee recorded paper and pencil tests for instrument panel reading and troubleshooting are shown in Exhibits 8 and 9.

Exhibit 7

Example of Scoring Checklist for Scoring Performance in Multiple Areas

PERFORMANCE EXAMINATION
WELDING
SCORING CHECKLIST

Time Started ____________

TOOLS AND MATERIALS

1. Holds torch at 45° angle to work except for start or finish 1
2. Always concentrates flame on base metal, not on rod 1
3. Uses proper flux consistency (free flowing flux) 1
4. Selects proper size rod for given metal thickness 1
5. Selects proper size welding tip for given metal thickness 1
6. Restricts cleaning of base metal to width of weld 1

PROCEDURE

7. Examines metal for dirt or grease. Cleans both metal and rods 1
8. Sets metal on jigs. Mixes flux. Fluxes both base metal and rods 1
9. Adjusts oxygen acetylene regulators to 3-8 pounds (no credit if pressure on oxygen does not equal pressure on acetylene) 1
10. Lights torch and adjusts to slightly carborizing flame (feather should be no more than 1½ times inner cone) 1
11. Pre-heats base metal 1
12. Tacks metal from center to each end, or from center to each end alternately; tacks 1¼ to 1½ inches apart 1
13. Welds from center to one end 1
14. Reverses metal and welds from center to other end 1
15. Uses correct torch and rod motions while welding 1
16. Dips and washes making sure that all flux is removed 1
17. Time finished ____________. Finished in 17 minutes or less 1

SAFETY

18. Shirt neck and sleeves buttoned 1
19. Makes sure fire extinguisher in area before igniting torch 1
20. Makes sure that gas bottles are in an upright position 1
21. Uses friction lighter to ignite torch and holds lighter on bench when igniting torch ... 1
22. Does not open acetylene cylinder valve more than 1½ turns 1
23. Uses goggles when welding 1

MEASUREMENT OF THE FINAL PRODUCT

24. Start of weld uniform with rest of weld 3
25. End of weld uniform with rest of weld 3
26. Uniform penetration for entire first 3 inches 3
27. Uniform penetration for entire last 3 inches 3
28. Bead width 3-5 times metal thickness for entire first 3 inches 3
29. Bead width 3-5 times metal thickness for entire last 3 inches 3
30. Bead height 26-50% of thickness for entire first 3 inches 3
31. Bead height 25-50% of thickness for entire last 3 inches 3
32. No bead irregularity in entire first 3 inches 3
33. No bead irregularity in entire last 3 inches 3

Total Score ____________

Exhibit 8

Example of Examinee Recorded Instrument Reading Test

IN COLUMN I PLACE THE NUMERICAL READING FOUND ON THE INSTRUMENT.

IN COLUMN II INDICATE WITH A "Yes" OR "No" WHETHER OR NOT THE READING IS WITHIN THE NORMAL OPERATING RANGE.

IN COLUMNS III AND IV GIVE THE MAXIMUM AND MINIMUM READINGS AT WHICH IT WOULD STILL BE SAFE TO OPERATE THE ENGINE.

TIME LIMIT: TEN MINUTES.

Instrument Number	I Reading	II Within Norm. Range (Yes or No)	III Maximum Safe	IV Minimum Safe
1.				
2.				
3.				
4.a				
4.b				
5.				

Exhibit 9

Example of Examinee Recorded Trouble Shooting Test

CONTROLLER TROUBLE SHOOTING

DIRECTIONS: Each controller is energized. There is a trouble in one of the circuits of each controller. At each controller, push the start button and observe the action of the motor and/or controller. Using the information you get from this operation and the equipment provided locate the troubles in each of the controllers. Indicate your answers in the table below by placing an (X) under the controller number and opposite the condition or part which is causing the trouble.

FOR EXAMPLE: If you found a trouble in the start button of controller "B" you place an (X) as shown below:

	Controller Number				
PILOT CIRCUIT	A	B	C	D	E
a. Control fuses					
b. Stop buttons					
c. Start buttons		x			

The main thing is to find the troubles. Don't waste time, extra credit will be given for rapid work, but observe all safety precautions. Are there any questions?

	CONTROLLER NUMBER									
POWER SUPPLY CIRCUIT	1	2	3	4	5	6	7	8	9	10
a. Blown fuses										
b. Open circuit breaker or switch										
c. Poor connection in circuit										
PILOT CIRCUIT (Control Circuit)	1	2	3	4	5	6	7	8	9	10
a. Control fuses										
b Stop buttons										
c. Start buttons										
d Auxiliary contacts										
e. Overload relay contacts										
f. Poor connections										
MAIN LOAD CIRCUIT (Motor Load Circuit)	1	2	3	4	5	6	7	8	9	10
a. Main line contacts										
b. Overload relay elements										
c. Poor connections										

KEEP YOUR ANSWER SHEET FACE DOWN WHEN NOT WRITING

Pictorial Performance Tests

A variant of the examinee recorded performance test is that in which the testee has to examine a picture or diagram in which some facet of the training or job is being depicted incorrectly. The examinee's task is to write down or tell what is wrong with the picture. An example of a pictorial performance test item measuring safety knowledge is shown in Exhibit 10. In this example, an airman is discharging a capacitor with a screw driver rather than with the proper tool for this purpose.

A prerequisite to the construction of a performance test is a behavioral job analysis. A behavioral job analysis determines the behaviors required for task performance; the importance of each task element, and an indication of how to tell whether one has correctly performed each task element.

The advantages of performance tests are: (1) realism, (2) practicality, (3) objectivity, (4) content validity, and (5) freedom from verbal requirements.

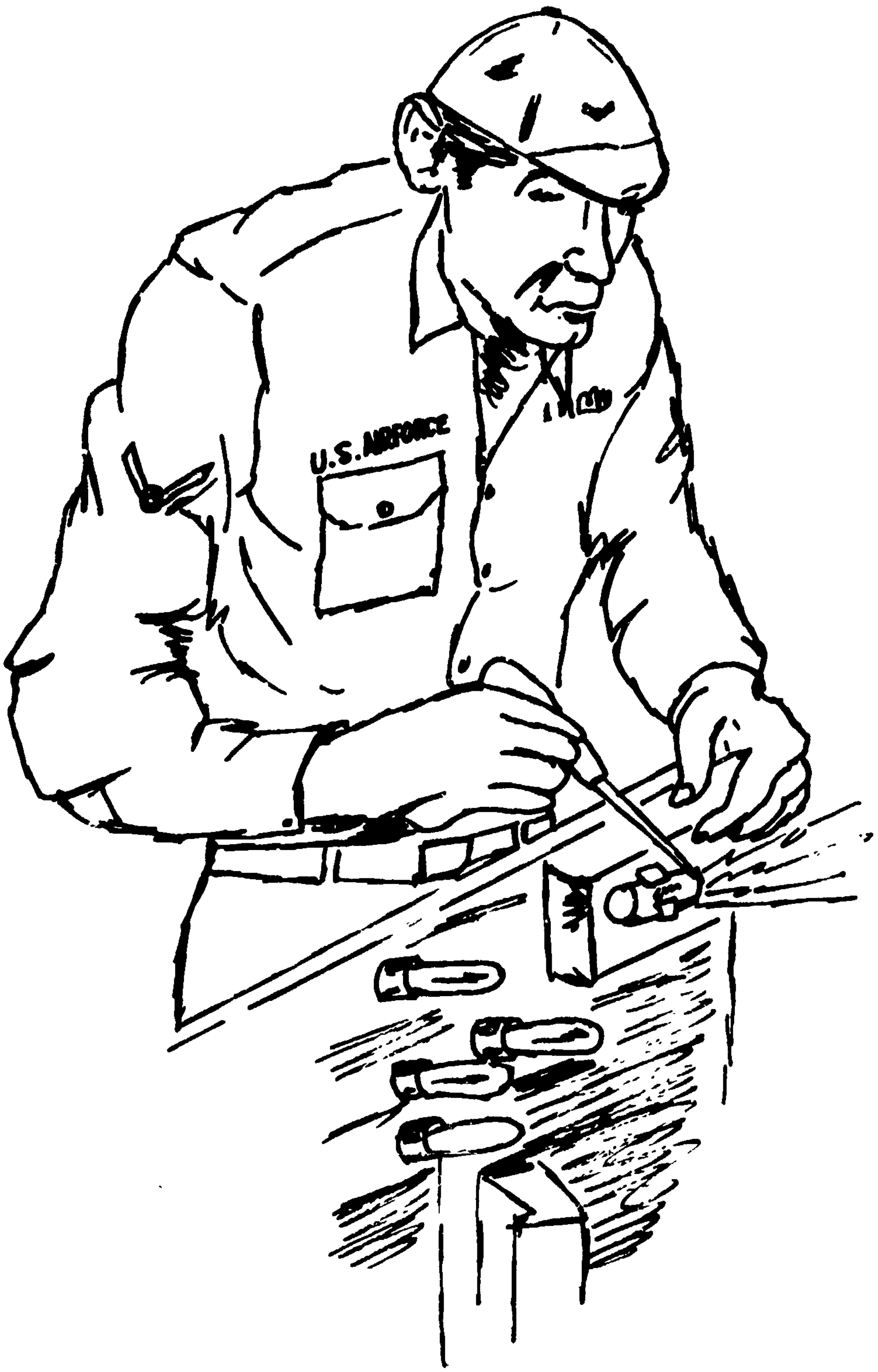
One disadvantage of performance tests is their potentially high cost. Often the actual on-the-job equipment or apparatus is needed for the test purposes. Great care should be taken in determining the degree of fidelity actually needed for performance test purposes.

Performance tests are also costly to administer. This is especially true if you can test only one man at a time, because there is only one piece of apparatus available.

Many performance tests require individual administration, i.e., one examiner for each examinee. This places a serious strain on available test administrative manpower. These very points were supported by the student measurement specialists at the various Air Force technical training centers who, when interviewed, claimed that performance tests (however practical and appropriate) presented serious administration difficulties.

Exhibit 10

Example Item from a Pictorial Performance Test



Proving Grounds

Perhaps the most sophisticated type of performance test is the personnel proving ground. In the proving ground, the trainee is placed on the job. An attempt is made to cycle him through all the job tasks in a short period of time. As he performs each job, the trainee is evaluated and he, in turn, evaluates the training he received, in relation to the job.

Ratings and Related Methods

Supervisory Ratings

Supervisory ratings, although widely used, are one of the most unreliable, biased, and contaminated methods for evaluating performance. Several factors which can contribute to poor or inadequate ratings are: friendship, quick guessing, jumping to conclusions, first impressions, prejudice, halo, errors of central tendency, and errors of leniency. Of these, the last three are the most important. Halo exists when a rater allows his overall, general impression of a man to influence his judgment of each separate trait on the rating scale. You can determine if halo exists in the responses to a rating scale by intercorrelating all of the items. If there is a moderate or high correlation among most of the items, then halo probably exists. A more precise method for determining halo is given by Guilford.* The method depends on multiple raters and the use of a somewhat sophisticated statistical analytic technique called analysis of variance.

Errors of leniency occur when a rater tends to use only the upper portion of the rating scale when rating all or most of his men. Errors of central tendency occur when the rater uses only the middle portion of the rating scale when rating his men. Considerable evidence exists which demonstrates that rater training can reduce these sources of bias so that the resultant ratings are, at least, minimally useful.

*Guilford, J. P. Psychometric methods. New York: McGraw-Hill, 1954. Pp. 281-283.

Generally, ratings improve to the extent that we prevent the influence of the rater's own idiosyncrasies from affecting his observation of subordinate behavior. The evaluator must observe and record behavior in objective terms. If this suggestion seems mechanistic, it is meant to be that way. The more that a rater can become like a behavioral metering device, the less likely that he will contaminate the evaluation. Also, it will help immensely if the items are couched in behavioral rather than in relative or evaluative (e.g., superior, above average, etc.) terms.

In general, ratings are, at best, a haphazard method for evaluating training, performance, student achievement, or job behavior. Their only real advantages are speed and ease of construction.

If other more objective methods are feasible, then they should be used.

Ranking

Ranking is a rating variant which controls for the errors of leniency and central tendency. In the classical *pair comparison method*, every person being rated is separately compared with every other person on each attribute under consideration. The easiest way to accomplish this is to put each trainee's name on a separate index card, along with an identification number. If you have 15 trainees to be ranked, you will have 15 numbered index cards. Next, you construct a matrix with each person's name across the top and down the side. Then, take your first card and place it face up in front of you. This card is known as the "standard." Then, take the second card and compare, on the attribute under consideration, that man with the man represented by the standard card. When you decide which is superior, you place either a *1* (if the standard man is superior), or a *2* (if the comparison is superior) in that part of the matrix where the standard and he comparison intersect. Next, lay card number two aside and perform the same procedure with card number three and the standard. Continue doing this until the standard is compared with all the remaining cards. When you finish, lay the standard aside. It is not to be used again. Now, card number two will be your standard, and it will be compared with all the remaining cards. Set card number two aside, shuffle the remaining cards and resume the comparison procedure, as above. When all the cards have been compared with the second standard, choose a third standard and continue the procedure until all cards, except the last, have served as standard. To treat the data, simply count up the number of times each man was preferred, and enter that number in a column to the right of the matrix. Next, you rank each person on the basis of the number of times he was preferred. The end result will appear as shown in Exhibit 11.

Exhibit 11

Example of Data Treatment for Pair Comparison Rating

	Man 1	Man 2	Man 3	Man 4	Man 5	Number of Times Preferred	Rank	Standard Score
Man 1	-	1	1	4	5	2	3	50
Man 2			2	4	5	1	4	43
Man 3				4	5	0	5	25
Man 4					5	3	2	57
Man 5						4	1	75

A distribution of ranks will be the result of this procedure. A rank distribution, though, does not accurately represent the behavior of all the men being evaluated. To correct for this inadequacy, the rank distribution should be transformed to a normal distribution. A normal distribution will more accurately represent the behavior of the men being evaluated. In order to convert the rank distribution to a normal distribution, you must use the table presented as Exhibit 12. The entries in this table are standard scores.* To obtain the standard score for each man being evaluated, you first locate, in the left hand column of the table, the number of times the man was preferred. You, then, look across the top column and find the number of persons you have evaluated. The entry in the table where these two numbers intersect is the standard score assigned to the man. For example, if the man you wish to evaluate was preferred 10 times and the group size was 16, then the standard score assigned to the man is 54. On the other hand, if the group size was 24, then the standard score assigned to the man would be 48. This particular example also demonstrates that the evaluator must know the group size as well as the man's rank in order to determine relative standing. The man who is ranked tenth in a group of 12 is certainly different from the man who is ranked tenth in a group of 25.

If each man in the evaluation group is ranked by more than one evaluator,** you must follow a certain procedure for combining ratings. An averaging procedure is used, but it is not the ranks that are averaged. The standard scores must be averaged--not the ranks.

*The standard score concept will be discussed more thoroughly in Chapter IV of this handbook.

**Multiple ıatings are desirable in most situations.

Exhibit 12

TABLE FOR CONVERTING PAIR COMPARISON DATA TO STANDARD SCORES

Number of Times Preferred	Number of Persons Rated																				
	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
0	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25	25
1	43	41	40	39	38	38	37	37	36	36	35	35	35	34	34	34	33	33	33	33	33
2	50	47	46	44	43	42	41	41	40	40	39	39	39	38	38	37	37	37	37	36	36
3	57	53	50	48	47	46	45	44	43	43	42	41	41	41	40	40	40	39	39	39	38
4	75	58	54	52	50	49	47	46	46	45	44	44	43	43	42	42	41	41	41	41	40
5		75	60	56	53	51	50	49	48	47	46	46	45	45	44	44	43	43	42	42	42
6			75	61	57	54	53	51	50	49	48	47	47	46	46	45	45	44	44	44	43
7				75	62	58	55	53	52	51	50	49	48	48	47	47	46	46	45	45	45
8					75	62	58	56	54	53	52	51	50	49	49	48	47	47	46	46	46
9						75	63	59	57	55	54	53	52	51	50	49	49	48	48	47	47
10							75	63	60	57	56	54	53	52	51	51	50	49	49	48	48
11								75	64	60	58	56	55	54	53	52	51	51	50	49	49
12									75	64	61	59	57	55	54	53	53	52	51	51	50
13										75	65	61	59	57	56	55	54	53	52	52	51
14											75	65	62	59	58	56	55	54	53	53	52
15												75	66	62	60	58	57	56	55	54	53
16													75	66	62	60	58	57	56	55	54
17														75	66	63	60	59	57	56	55
18															75	66	63	61	59	58	57
19																75	66	63	61	59	58
20																	75	67	63	61	60
21																		75	67	64	62
22																			75	67	64
23																				75	67
24																					75
25																					

When there are 10 or more instructors evaluating a group of students, another, more sophisticated, combinatory procedure can be used. In this procedure, you must first calculate the number of times each individual is judged more favorable than every other individual by the group of raters. This approach is facilitated by the use of a matrix. A sample frequency matrix is presented as Exhibit 13. Here, five students were evaluated by 12 instructors. In this matrix, for instance, man 3 was preferred over man 1 eight times, while man 1 was preferred over man 3 only four times.

Exhibit 13

Example of Initial Frequency Matrix for Pair Comparison Data When More than Ten Raters are Involved

	Man 1	Man 2	Man 3	Man 4	Man 5
Man 1	6	5	8	7	6
Man 2	7	6	9	7	8
Man 3	4	3	6	2	7
Man 4	5	5	10	6	4
Man 5	6	4	5	8	6

These values are then converted to proportions (of 12). The sample proportion matrix corresponding to the frequency matrix (Exhibit 13) is presented as Exhibit 14.

Exhibit 14

Example of Proportion Matrix for Pair Comparison Data

	Man 1	Man 2	Man 3	Man 4	Man 5
Man 1	.50	.42	.67	.58	.50
Man 2	.58	.50	.75	.58	.67
Man 3	.33	.25	.50	.17	.58
Man 4	.42	.42	.83	.50	.33
Man 5	.50	.33	.42	.67	.50

As in the previous example, these proportions must be converted to standard scores. Conversion of the above proportion matrix using Exhibit 16 is presented in Exhibit 15. The bottom row of Figure 15 presents the final scale values.

Exhibit 15

Example of Standard Score Matrix for Pair Comparison Data

	Man 1	Man 2	Man 3	Man 4	Man 5
Man 1	.00	-.20	.47	.20	.00
Man 2	.20	.00	.67	.20	.47
Man 3	-.44	-.67	.00	-.95	.20
Man 4	-.20	-.20	.95	.00	-.44
Man 5	00	-.44	-.20	.47	.00
Sum	-.44	-1.51	1.89	-.08	.23
Mean (Scale Value)	-.09	-.30	.38	-.02	.05

Exhibit 16

Standard Score Values for Each Proportion from .01 to .99

Proportion	Standard Score	Proportion	Standard Score	Proportion	Standard Score
.99	2.33	.66	.41	.33	-.44
.98	2.05	.65	.39	.32	-.47
.97	1.88	.64	.36	.31	-.50
.96	1.75	.63	.33	.30	-.52
.95	1.65	.62	.31	.29	-.55
.94	1.56	.61	.28	.28	-.58
.93	1.48	.60	.25	.27	-.61
.92	1.41	.59	.23	.26	-.64
.91	1.34	.58	.20	.25	-.67
.90	1.28	.57	.18	.24	-.71
.89	1.23	.56	.15	.23	-.74
.88	1.18	.55	.13	.22	-.77
.87	1.13	.54	.10	.21	-.81
.86	1.08	.53	.08	.20	-.84
.85	1.04	.52	.05	.19	-.88
.84	.99	.51	.03	.18	-.92
.83	.95	.50	.00	.17	-.95
.82	.92	.49	-.03	.16	-.99
.81	.88	.48	-.05	.15	-1.04
.80	.84	.47	-.08	.14	-1.08
.79	.81	.46	-.10	.13	-1.13
.78	.77	.45	-.13	.12	-1.18
.77	.74	.44	-.15	.11	-1.23
.76	.71	.43	-.18	.10	-1.28
.75	.67	.42	-.20	.09	-1.34
.74	.64	.41	-.23	.08	-1.41
.73	.61	.40	-.25	.07	-1.48
.72	.58	.39	-.28	.06	-1.56
.71	.55	.38	-.31	.05	-1.65
.70	.52	.37	-.33	.04	-1.75
.69	.50	.36	-.36	.03	-1.88
.68	.47	.35	-.39	.02	-2.05
.67	.44	.34	-.41	.01	-2.33

From these data, we can ascertain that man three was judged most favorable by the raters and that man two was judged least favorable.

Fractionation Method

One rather sophisticated rating method combines the pair comparison technique with the magnitude estimation. In the fractionation method, one divides 100 points between two stimuli. For example, in comparing the size of two objects, the assignment of 75 points to one member of a pair and 25 points to the other member of the pair would indicate that the first is three times the size of the second. One can, of course, apply this technique to the evaluation of individual traits or performance of students. If you are evaluating two individuals and you assign one a value of 80 and the other a value of 20, this means that the first individual possesses four times as much of the trait being evaluated as the second. A 50-50 split means that both individuals perform about the same with regard to the trait in question. In order to illustrate the method, a modification of an example given by Comrey* will be presented. Comrey had 47 persons judge five stimuli using the fractionation technique. A matrix showing the average number of points assigned to each stimulus is given in the upper half of Exhibit 17.

The next step in the procedure is to convert each of the entries in the upper half of Exhibit 17 to its corresponding logarithm. These logarithmic data are shown in the lower half of Exhibit 17.

*Comrey, A. L. A proposed method for absolute ratio scaling. Psychometrika, 1950, 15, 317-325.

Exhibit 17

Initial Steps in Treatment of Fractionation Data

	Average Number of Points Assigned to Each Stimulus				
Stimulus	1	3	5	7	9
1	-	44.91	36.09	28.74	11.00
3	55.09	-	41.04	34.51	12.79
5	63.91	58.96	-	41.34	16.00
7	71.26	65.49	58.66	-	21.06
9	89.00	87.21	84.00	78.94	-

	Corresponding Logarithms for the Values Presented Above				
Stimulus	1	3	5	7	9
1	-	1.652	1.557	1.458	1.041
3	1.741	-	1.613	1.538	1.107
5	1.806	1.771	-	1.616	1.204
7	1.853	1.816	1.768	-	1.323
9	1.949	1.941	1.924	1.897	-

Then, the values for each of the corresponding pairs are subtracted. For example, you first subtract 1.741 from 1.652, resulting in a value of -.089. The next subtraction would be 1.806 from 1.557, resulting in -.248. This procedure is continued until every pair is subtracted. The results for the entire matrix are shown in Exhibit 18. Also shown in Exhibit 18 are the sums, means, and antilogs. These antilogs represent the scale values of the stimuli on the characteristic under consideration.

Exhibit 18

Subtracted Log Values and Physical Scale Values of the Stimuli in Inches

Stimuli	1	3	5	7	9
1	0	-.089	-.248	-.394	-.908
3	.089	0	-.157	-.278	-.834
5	.248	.157	0	-.152	-.720
7	.394	.278	.152	0	-.574
9	.908	.834	.720	.574	0
Sum	1.639	1.181	.467	-.251	-3.034
Mean	.328	.236	.093	-.050	-.607
Antilog	2.127	1.722	1.240	.891	.247

Order of Merit

In the pair comparison and the fractionation methods, as the number of trainees increases, the number of comparisons increases geometrically. For example, when you have 15 trainees, there are 105 comparisons. But, with 30 trainees, there are 435 comparisons. Although it is more precise than rating scales, use of the pair comparison method is not recommended if you have to judge any more than 20 or 25 persons.

A ranking method which also controls for the leniency and central tendency errors and which is much simpler to use than pair comparisons is called the order of merit method. In this method, you just arrange your set of index cards containing the names of the persons to be rated on a particular trait in order from best to worst.

A preferred variation of this method is known as alternation ranking. In this procedure, you first list the names of the persons to be ranked on the left side of a piece of paper. Scan the list and select the best. His name is written at the top of the right side of the page and crossed off the left side of the paper. Then, the list is scanned again, and the worst man is selected from the list. His name is put on the bottom of the right hand side of the page and crossed off the list on the left. This procedure is continued until all the names have been crossed off the list. A partially completed visual presentation of this method is presented in Exhibit 19.

Exhibit 19

Example of Alternation Ranking Procedure

Person to be Ranked	Ranking from Best to Worst	
John Jones	Mark Davis	1
Joseph Doakes	Jessie Goldman	2
~~John Doe~~		3
Tom Smith		4
~~Mark Davis~~		5
Tyrone Sullivan		6
~~Jessie Goldman~~		7
Peter Thomas		8
Mario Pilosi		9
~~Ivan Ivanovich~~	Ivan Ivanovich	10
Robert Watkins	John Doe	11

It is obvious that the order of merit method of ranking is more flexible than pair comparisons. It can be employed to evaluate many more people and many more traits in a shorter amount of time. It possesses the added advantage over ratings of being able to control for the leniency and the central tendency errors. Halo error, however, can still occur, especially if the men are ranked on more than one trait.

Other Evaluation Methods

Gain Scores and Final Exam Grades

Many persons argue that gain scores are superior to final exam grades for both training evaluative and student achievement measurement purposes. This is because gain scores are held to control for initial level of knowledge. A gain score is the difference between the student's knowledge before and after training. Comparable forms of a test are usually given before and after training to derive the difference score. The problem in the before and after measurement of gain is that when small significant increases are registered, there may actually be a tremendously large increase in knowledge. This paradoxical result comes from the inequality of measurement at different points along the scale. Actually, knowledge increases faster than test scores. One can rarely find a significant positive correlation between initial test scores and gain scores (often there is an inverse correlation). This is contrary to expectation, since it is expected that the more intelligent student will learn more and that the more interested student will be motivated to study more. One can partially explain the anomaly on the basis that students who already know a lot do not have much left to learn.

Another related problem is the ceiling effect. This occurs when the initially bright student already has most of the items on the pretest correct and, accordingly, doesn't have much room for improvement on the posttest.

A related approach involves separating the initially bright from the initially dull students. This is done to correct for the loss of time by the initially high scoring student who has to waste time while others are learning low level material. It is possible that if the bright student's instruction started off at a higher level, his gain will be greater.

Rate of Gain

In the rate of gain method, one controls for variables (using the part correlation technique) which can affect the traditional pre- and posttest gain scores. Through this technique, one might, for instance, remove the effects of intelligence from both the pre- and posttest scores. The residual difference, with intelligence controlled, is pure "rate of gain" and is held to be one of the more sensitive criteria of achievement.

Task Analytic Evaluation

In one sense, training for a job may be considered to be acceptable if the average trainee performs with proficiency on the highly important tasks of a job. The training may be considered to be faulty if the average worker performs poorly on the very important tasks. Obviously, this conceptual method of evaluation involves relating job analytic information (importance) to average job behavior. The method points to deficiencies in the program which need emphasis, and parts of the program which need deemphasis. A technique developed by Siegel and Schultz is presented as an example. Siegel and Schultz* suggest a matrix approach. First the average proficiency is calculated for a representative group of men on each task. Next, the importance of each task is determined. Then, each task is sorted into one of the nine cells shown in Exhibit 20. Cell A consists of tasks which are highly important and on which the average worker is highly proficient. Cell F, though, is composed of tasks which are relatively low in importance, but on which the average worker is moderately proficient. This matrix can be studied from several points of view. A training index, an overtraining index, and an undertraining index can be derived from the matrix. The training index is used to determine if training has been effective. In order to quantify this approach, let each task in cells A, E, and J be given a weight of 2, each task in cells B, D, F, and H a weight of 1, and the tasks in cells C and G a weight of 0. Tasks which meet the ideal are thus scored 2; those which depart from the ideal are scored 1 or 0, depending on the extent to which

*Siegel, A. I., & Schultz, D. G. Evaluating the effects of training. Journal of the American Society of Training Directors, 1961.

they deviate. The training index is computed by summing the weights for all the tasks in the job and dividing this number by twice the number of tasks involved. This procedure will yield an index which varies between 0 and 1. Perfect training is the condition in which all the tasks fall into cells A, E, and J. In this instance, the training index is 1. If all the tasks were to fall into cells G and C, then the training index would be 0.

Exhibit 20

Matrix for Classifying Job Tasks by Importance (I) and Average Proficiency (P)

		I		
		High	Moderate	Low
	High	A	B	C
P	Moderate	D	E	F
	Low	G	H	J

Overtraining is defined as the extent to which average proficiency on various tasks exceeds the level indicated by their importance. If the average graduate is highly proficient on tasks of little importance, overtraining exists. The overtraining index is calculated by assigning a weight of 2 to all tasks in cell C, a weight of 1 to tasks classified in cells B and F, and a weight of 0 to tasks classified in the remaining cells. To obtain the overtraining index, the sum of the weights for all tasks in the job is divided by twice the number of tasks involved. The closer this index is to 1, the more certain we can be that overtraining exists.

In undertraining, we are interested in the extent to which the moderately and highly important tasks are performed poorly. If a weight of 2 is assigned to cell G, a weight of 1 to cells D and H, and a weight of 0 to the remaining cells, the undertraining index can be obtained. This is the sum of the weights for all tasks according to the cell in which they are classified, divided by twice the number of tasks. In this instance, undertraining increases as the index approaches 1. An index of 0 in this case suggests suitable training.

The three indices described above each shed a different light on the training program. When employed compositely the indices can provide the basis for a rather thorough understanding of the effectiveness of the training program. The training index tells the user whether the training program is generally effective in the sense that the graduates are able to perform the important tasks on the job. This index reflects the extent to which the graduates approach the paradigm of high on-the-job proficiency for relatively important tasks and low on-the-job proficiency for relatively unimportant tasks. If training time is limited, it could be maintained that a trainee ought, first of all, to learn how to do the important aspects of the job. One might wish or demand that the new worker be highly proficient on all the tasks of the job. But this may be impractical from a realistic standpoint and inefficient from the standpoint of minimizing the time and cost of training.

The overtraining index emerges directly from the concept inherent in the training index. In terms of the training index, the training product is not considered suitable if the graduate is very proficient on unimportant tasks. Objection may be raised to "marking down" a training program because it produces this kind of proficiency. It is certainly true that overtraining in this sense is preferable to undertraining, i.e., producing a graduate who is not proficient on tasks which are important. The overtraining index yields insight into the overtraining aspect of the training graduates' job suitability. This index indicates the extent to which the graduates are highly proficient, even on the unimportant tasks of the job.

To complete the picture of the meaning of the proficiency-importance matrix, the undertraining index quantifies the extent to which the training gives rise to workers who do not perform adequately the more important tasks of the job. Since it seems safe to assume that satisfactory job performance would ordinarily require the ability to carry out the more vital functions involved, this index places emphasis on any serious inadequacies of a training effort.

Questionnaire Measurement

The measurement of student opinions through student completed questionnaires is often an integral component of training evaluation. It makes very little sense to measure posttraining performance or achievement without measuring student attitudes and opinions as well. A student's posttraining performance will often be affected by his attitudes and opinions. A properly phrased questionnaire, completed by the students, can point to areas of deficiency in the training program and the student environment overlooked by the instructors and not easily identified through the other methods described in this manual.

Several areas can be inquired into through questionnaire methods, depending on the purposes of the training evaluator. One area is concerned with the affective or emotional reactions of the student to the training program. A second is concerned with specific thoughts and suggestions that the student may have regarding the formal or pedagogical aspects of a training program. A third area of inquiry may involve the classroom and laboratory environment. Different types of questionnaire items may have to be constructed in order to obtain these kinds of information.

There are some very complex questionnaire techniques. The methods for constructing complex questionnaire items are beyond the scope of this manual. We shall confine ourselves to a discussion of four of the simpler item forms.

Likert* type items simply ask the trainee to indicate the extent of his agreement or disagreement with a statement about the training program. Some sample Likert items, concerned with motivation and enjoyment of a training program, are presented in Exhibit 21.

*Rensis Likert, an industrial psychologist, was one of the first to construct and use this type of item.

Exhibit 21

Example of Likert Type Items

Directions

Please indicate whether you Strongly Agree (SA), Agree (A), Mildly Agree (MA), Mildly Disagree (MD), Disagree (D), or Strongly Disagree (SD) with each of the following statements (circle one).

---

1. This training program made the best possible use of training aids (circle one) SA A MA MD D SD

2. The instructor frequently allowed time for class discussion (circle one) SA A MA MD D SD

3. The tests were based on the material taught in class (circle one) SA A MA MD D SD

4. There was ample opportunity for review at the end of each class meeting (circle one) SA A MA MD D SD

5. The material was presented at the proper pace; neither too fast nor too slow (circle one) SA A MA MD D SD

6. There was a good balance between material and demonstration (circle one) SA A MA MD D SD

You will notice that the items in Exhibit 21 have six possible choices. The writer of these items forced the respondent to agree, or disagree, by not including a middle category. Sometimes, you may wish to supplant the MA and MD choices with "no opinion" or "indifferent" choices. However, good technical reasons exist for avoiding this, where possible. Likert type items can be designed to obtain both affective reactions and information from the students.

Semantic Differential

The semantic differential method uses bipolar adjectives to define the endpoints of semantic dimensions. To develop such a scale, the evaluator first must postulate a region of semantic space; then, several bipolar semantic scales are constructed whose axes pass through the center of this space. One should identify as many dimensions or axes as possible so that the dimensionality of the space is exhausted. With regard to training evaluation, semantic differential items require the student to indicate his feelings about various aspects of the training program. Semantic differential items can only be used to secure affective reactions about the training program from the trainees. Exhibit 22 presents sample items from a questionnaire based on the semantic differential technique.

Exhibit 22

Example of Semantic Differential Items

Directions

Place a checkmark (√) anywhere along the line between each pair of words to indicate your feelings about the training program.

Bad	└──┴──┴──┴──┴──┴──┴──┴──┘	Good
Interesting	└──┴──┴──┴──┴──┴──┴──┴──┘	Dull
Adequate	└──┴──┴──┴──┴──┴──┴──┴──┘	Inadequate
Hard	└──┴──┴──┴──┴──┴──┴──┴──┘	Easy
Too Complex	└──┴──┴──┴──┴──┴──┴──┴──┘	Too Simple
Ineffective	└──┴──┴──┴──┴──┴──┴──┴──┘	Effective

Sentence Completion

Sentence completion questions present the student with an item which he is required to complete. Generally, the stem or premise serves to channel the respondent's thoughts in a given direction in accordance with the objectives of the evaluator. Some examples of sentence completion questions are presented in Exhibit 23.

Exhibit 23

Example of Sentence Completion

Directions

Please complete each item as honestly and completely as you can. Your answers will be used to help revise and improve the training program.

1. The best aspects of this training program were ______________
____
____

2. The worst aspects of this training program were______________
____
____

3. The most needed improvements in this training program ________
____
____

4. The instructional methods ______________________________
____
____

5. In regard to the training facilities, I think that ______________
____
____

The open-ended question elicits a free response, one which is not restricted to any predetermined categories. The respondent replies in a natural and unstructured manner and in whatever terms or frame of reference he chooses. For this reason, the sentence completion type of question could lead to greater comprehensiveness and result in responses of higher validity than questions with a fixed set of responses. With a fixed set of responses, a superficial similarity may be enforced and the true response of certain respondents may not become known. Also, a specific question may be perceived differently by different people, and a difference in an answer may reflect the difference in interpretation. This kind of situation could be eliminated with the sentence completion question.

The disadvantages of the sentence completion item lie in its scoring. This type of question does not lend itself to precoding. Therefore, the tabulation of the results is considerably more involved than for a question with the fixed set of responses. One method for treating such data is to construct response categories and to tally the number of statements falling into each. One way to avoid this predicament will be reviewed in the next section.

Multiple Choice Questions

Multiple choice questions avoid the data treatment problems inherent in sentence completion items, since the response categories are already structured for the student. However, the open ended format must sometimes be employed as a precursor to the multiple choice format. This approach allows the derivation of the necessary multiple choice response categories. While the multiple choice response format is easier than the open ended format for the person completing the questionnaire, the multiple choice format may tend to lose data. Since the questionnaire constructor will probably include as options only those response categories which are most popular, interesting but outlying responses may be lost. Some examples of multiple choice questions are presented in Exhibit 24.

Exhibit 24

Example of Multiple Choice Questions

Directions

Please select the one answer which best reflects your feelings from among the choices for each of the following questions.

---

1. Check (✓) the one lesson in the flight training program which was the easiest for you to learn.

   a. Maneuvers ______
   b. Instrument Flight ______
   c. Navigation ______
   d. Radio Procedures ______
   e. Formation Flight ______
   f. Nomenclature ______

2. Check (✓) the one lesson in the pilot training program which was the most difficult for you to learn.

   a. Maneuvers ______
   b. Instrument Flight ______
   c. Navigation ______
   d. Radio Procedures ______
   e. Formation Flight ______
   f. Nomenclature ______

3. Check (✓) the one lesson which most needs improved training aids.

   a. Maneuvers ______
   b. Instrument Flight ______
   c. Navigation ______
   d. Radio Procedures ______
   e. Formation Flight ______
   f. Nomenclature ______

Exhibit 24 (cont.)

4. Check (√) the one improvement that you think is most needed in this program.

a. Better instructors ________
b. Better facilities ________
c. More leisure time ________
d. Better reference library ________
e. Less stress on academic and
more on practical aspects ________

5. Check (√) the one aspect in the training program you liked least.

a. Maneuvers ________
b. Instrument Flight ________
c. Navigation ________
d. Radio Procedures ________
e. Formation Flight ________
f. Nomenclature ________

Forced-Choice Items

As one would suspect a forced-choice items is an item where the examinee is forced to choose one of a number of alternatives. This format is often useful for purpose of rating oneself or another on a trait or characteristic. Although forced-choice items are often in the vein of "Are you still beating your wife" and therefore can cause examinee resistance, they are held to eliminate much response faking and dishonesty. In constructing such items care must be taken to equate the social desirability of the response alternatives. This is done to avoid having examinees respond to the desirability of an item only, rather than to the content of the item.

This questionnaire method, though, poses many more developmental problems than the methods previously presented. In the first place, forced-choice questions are very difficult to construct. The items first have to be administered to a sample of persons so that the social desirability of the items can be assessed. Second, factor analytic studies must be performed so that the items can be empirically grouped by the trait or belief being measured. Third, different items, of equal desirability, must be combined in a pair comparison framework in order to make up the questions.

Examples of forced-choice items where two undesirable and two desirable statements occur within one question are shown in Exhibit 25. Each sample provides three scoring levels. The student is given a +1 both for M responses to positively phrased statements and for L responses to negatively phrased statements. The student is assigned a -1 both for L responses to positively phrased statements and for M responses to negatively phrased statements. Response choices that are not checked with an M or an L are assigned a score of 0. Each of the items is related to one of the dimensions of interest. All of the positive and negative scores for each training dimension are, then, added up. For example, suppose the student completes a 20 item forced-choice questionnaire with instructor behavior as one of the training dimensions. Suppose the student's scores on the instructor behavior dimension were +9 and -3[*]. His total score, then, for instructor behavior would be +6. This procedure allows one to rank each relevant aspect of training so that areas of improvement can be delineated.

[*] Eight of the 20 instructor behavior items were not marked with either an M or an L.

Exhibit 25

Examples of Forced-Choice Items

Directions

Place an M next to the option in each set which best describes your reaction to the set. Place an L next to the option which most poorly or least describes your reaction. You can mark only two of the four options in each set.

1. a. The instructors were considerate of student needs □
   b. The course was at an appropriate difficulty level □
   c. The textual material was too difficult □
   d. The workbook was disorganized □

2. a. The training aids helped my understanding □
   b. The classroom facilities were inadequate □
   c. The laboratory work was necessary □
   d. The instructors were unfriendly □

3. a. The instruction was disorganized □
   b. The tests were fair □
   c. The material was clearly presented □
   d. The instructors failed to provide individual help □

The Interview

The interview is one of the oldest and most widely used forms of evaluation that we have. Almost every evaluation process involves some interviewing. This is a paradox, since the interview represents an extremely weak process with little evidence to support its use, while, at the same time, there is a great deal of evidence against it. This is not to say, though, that the interview should be eliminated, because through a systematic approach, it can be improved.

The most important purpose of the interview, in the present context, is to obtain evaluative information about either the training program or the student. There are three qualitatively different ways of obtaining this kind of information. These are:

1. Unsystematic

   This type of interview is unplanned, haphazard, free associating, and unstructured. The interviewer may have a set of questions he has developed which he thinks will do a good job. Usually, no particular sequence is followed and in many instances the interviewer will talk 90 per cent of the time. Generally, specific trainee or program characteristics which need to be identified are not identified. The interviewer winds up with an impression and gives superficial reasons for his decision. Most of the interviewing done today is of this type. Needless to say, this approach has proven highly unreliable and invalid.

2. Standardized

   The chief criteria of this method are high reliability and low variability. It is characterized by a series of questions that are always asked by all interviewers in the same sequence with the same wording and the same scoring procedures. This technique tends to eliminate the interviewer as a source of response variability (error) in the applicant. This type of interview has been criticized on the grounds that it is nothing more than an oral test with the tendency to obtain "pure" information.

3. Systematic

In this method, the interviewer enters with a reservoir of questions with the possibility of different questions for different trainees. This approach stresses interviewer flexibility so that he can dig for the things he is looking for and also that he will systematically record and evaluate impressions. The reasons for recording impressions are:

1. at that time all the highlights of the interview are in front of you
2. so that comparisons can be made with others
3. so that comparisons can be made with what is needed on the job (validity)

This is a problem solving, structured approach concerned with reliability, validity, and practicality.

In the systematic interview, we must plan what we are going to measure in advance. Information should be either relevant to the training program or be predictive of job success. In the interview, we attempt to develop information which will allow us to measure characteristics which are not better measured by other techniques (e.g., tests, questionnaires, etc.). The first step in interview construction is to identify the information that is needed in order to evaluate the training program or the student. Then, the information that can best be uncovered by the interview is identified. A series of questions is then constructed around each characteristic or requirement. One way to arrive at questions is to have a leaderless group composed of all those who are interested in the evaluative results. This group composes a pool of questions (in some cases there may be only one question) for each area of interest. An example of how the end result might look for one area of interest is presented below.

Information Needed	Questions	Some Possible Answers
1. How to motivate students	1a. Can you motivate students?	1a. Yes
	1b. How can we influence students to do things which are unpleasant?	1b. Order them. Threaten them. Impress them of its importance.
	1c. Can you describe some techniques that will make students study more?	1c. Praise. Opportunity for advancement. Leave. Prestige titles.

As the reader will note, question 1a is rather weak, since it requires only a "yes" or a "no" answer, while the other questions are designed to elicit detailed responses from the interviewee. It should also be noted that some possible responses are given which can be assigned weights or scores on a five or a seven point scale. For example, if the respondent says, "I'll order them, " or "I'd threaten them, " he would obviously get a low score for that area, but on the other hand, if he says, "I'd praise him for those parts of the lesson he does well, " he'd get a higher score. This approach is perfectly suited to graphic rating form techniques. If the interviewer doesn't feel that he has obtained a true or complete picture of the interviewee's thoughts in the area, he can continue to query and probe him until a rating can be made. When this is completed, the interviewee is questioned on the next area of interest.

One can see that the procedure described has the advantage of tests in that it is objective and systematic and at the same time it is flexible in that the interviewer only has to ask as many questions as it takes to come to a decision in a specific area. When the interview is completed, the rating forms for the abilities measured are treated like test scores subsequent to statistical evaluation of the interview.

CHAPTER IV

QUANTITATIVE METHODS

Once the training evaluative data have been collected, they must be treated (statistically analyzed) in order that firm, conclusions may be drawn. Statistical analysis provides methods for summarizing the data and presenting them so that they can be interpreted by both you and others. These analyses also tell you how much confidence you can have in your data, i.e., whether any obtained differences or relationships are real or whether they can be attributed to chance.

This chapter is not meant to be exhaustive. It describes some of the various statistical techniques with which all evaluators should be familiar. Some of the simpler techniques will be explained in detail, since they can be accomplished easily on a desk calculator, while some of the more complicated methods will only be described. You will need the aid of a computer programmer, a computer, or a statistician to accomplish the more complicated statistical techniques.

Central Tendency

On may occasions you will be interested in obtaining measures of central tendency. There are three measures of central tendency: mean, median, and mode. The mean is, of course, the arithmetic average of a group of scores. It is obtained by adding together all the scores in your sample and dividing by the number of persons (N) in the sample. The notation for the mean of all raw scores is usually $\bar{x}$. A small x is usually used as the notation for a single raw score. The Greek letter sigma (Σ) indicates the arithmetic operation of addition. Your formula for computing the mean, then, reduces to:

$$\bar{x} = \frac{\Sigma x}{N}$$

The median is the midpoint or middle score of a set of scores when the scores are arranged from lowest to highest. The mode, on the other hand, is the most common or frequent score in a set of scores. You will have little occasion to use these latter two measures of central tendency, since most statistics require the use of means rather than medians or modes. Under special circumstances[*], though, a median is preferred to a mean. A hypothetical test score distribution is presented in Exhibit 26 with the mean, median[**], and mode calculated.

Exhibit 26

Sample Calculation of Mean, Median, and Mode

4	
5	
7	
8	
9	N = 13
10	
11	$\bar{x}$ = 10.92
12	
12	Median = 11
13	
15	Mode = 12
16	
20	
Σ 142	

[*] When the score distribution is highly skewed or distorted.

[**] When there is an even number of scores, the median is the average of the two middle scores.

Dispersion

Most statistical analyses require, in addition to the calculation of a measure of central tendency, the computation of dispersion or variability. The three measures to be discussed here are the range, the standard deviation, and the variance. The range is simply the distance between the highest and lowest scores of the distribution. This measure of dispersion is rarely used except as a very gross description of a distribution. The standard deviation (sometimes called sigma and denoted by σ) and the variance, which is the standard deviation squared, are the most commonly used measures of variability. Standard deviations, although they vary in size across distributions, have certain interesting properties which make them very useful. For example, the distance between the mean and the first standard deviation above the mean, in a normal distribution, always contains 34 per cent of the cases. This is also true of the distance between the mean and one standard deviation below the mean. The distance between the first and second standard deviation above the mean is occupied by about 13.5 per cent of the cases. Also, the distance between the second and third standard deviation contains about two per cent of the sample. These distances and percentages are pictured in Figure 1.

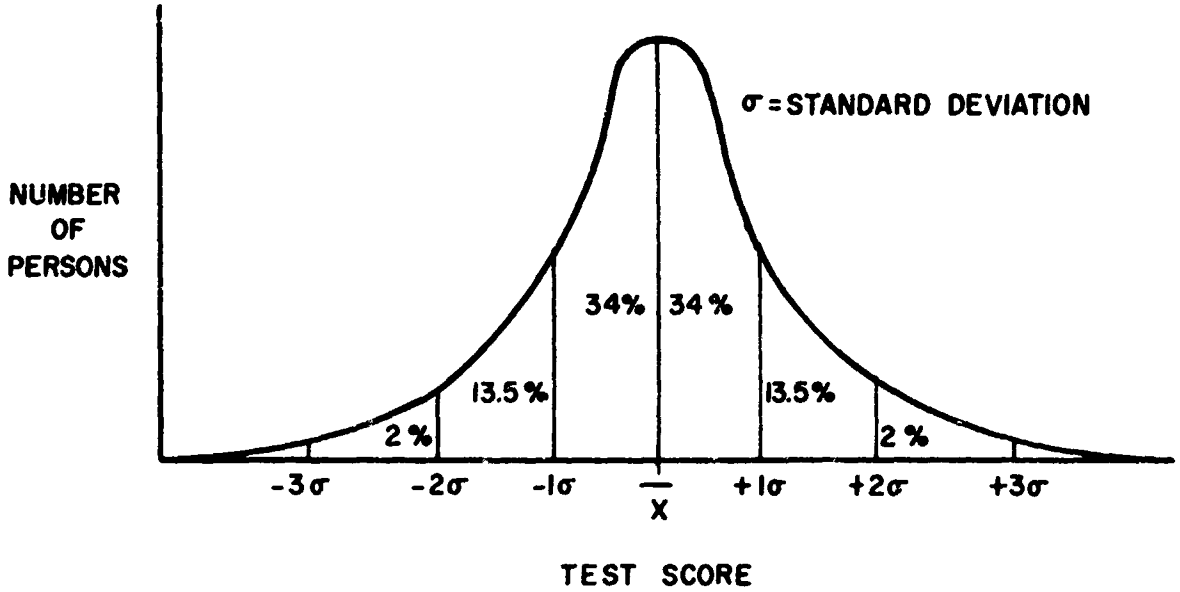


Figure I. Percentage of cases within various sigma limits.

You can see from Figure 1 that three standard deviations on each side of the mean will account for almost all of the cases in the test score distribution. The reason why we use the standard deviation in most of our calculations is because of its invariant properties across distributions. Typically, a distribution of scores is described fully by its mean and standard deviation.

The standard deviation is calculated by summing the deviations of each score from the mean, squaring them, and dividing by the number of cases. The square root of this number will give you the standard deviation. The formula for the standard deviation is:

$$\sigma = \sqrt{\frac{\Sigma(\bar{x}-x)^2}{N}}$$

With the same score distribution used earlier, the computation of a standard deviation is illustrated in Exhibit 27.

Exhibit 27

Sample Calculation of a Standard Deviation

x	$(\bar{x}-x)$	$(\bar{x}-x)^2$
20	+9.08	82.45
16	+5.08	25.81
15	+4.08	16.65
13	+2.08	4.33
12	+1.08	1.17
12	+1.08	1.17
11	+0.08	0.01
10	-0.92	0.85
9	-1.92	3.69
8	-2.92	8.53
7	-3.92	15.37
5	-5.92	35.05
4	-6.92	47.89
		Σ = 242.97

$$\sigma = \sqrt{\frac{242.97}{13}} = \sqrt{17.15*} = 4.14$$

* The variance of the above distribution is 17.15

Standard Scores

It is essential for the training evaluator to gain an understanding of the standard score concept. You may have encountered standard scores before reading this report, but they probably were referred to by another name. The results of many nationally administered tests are given in standard score format. The scores on the Scholastic Aptitude Tests of the College Entrance Examination Board are given in standard score form.

A standard score is always expressed in standard deviation terms. For example, if an individual's score on the arithmetic test was one standard deviation above the mean, at the 84th percentile, his standard score would be +1.00. Similarly, if this individual's score fell one standard deviation below the mean, at the 16th percentile, then his standard score would be - 1.00. If one's score was one-half of a standard deviation above the mean, at the 69th percentile, then his standard score would be +.50. In this way, one can assign a standard score which corresponds to every raw score in the distribution.

A standard score is computed by subtracting the raw score from the mean and dividing this value by the standard deviation. The standard scores for each raw score in our hypothetical distribution are shown in Exhibit 28.

As for the example shown in Exhibit 28, the mean of a standard score distribution is always 0.00 and the standard deviation of the standard score distribution is always 1.00. If the user wishes to avoid negative scores and decimals, he can multiply each obtained standard score by 10 and add 50 to the obtained value. This operation shifts the scale so that the mean is placed at 50 and the standard deviation is set equal to 10.

Exhibit 28

Sample Calculation of Standard Scores

		Standard Score
x	$(\bar{x}-x)$	$\bar{x}-x/\sigma$
20	+9.08	+2.19
16	+5.08	+1.23
15	+4.08	+0.99
13	+2.08	+0.50
12	+1.08	+0.26
12	+1.08	+0.26
11	+0.08	+0.02
10	-0.92	-0.22
9	-1.92	-0.46
7	-3.92	-0.95
5	-5.92	-1.43
4	-6.92	-1.67

Standard scores allow us to perform some mathematical manipulations which could not otherwise be performed. For example, suppose you wish to determine an overall score for an individual who has taken two tests. Suppose, further, that the standard deviations of the two tests are 7.84 and 4.14, respectively. If you want each test to count equally in the determination of final grade, you cannot add the scores of the two tests together. You cannot add them together, because the test with the higher standard deviation would contribute more variation to the final score than the test with the smaller standard deviation. Actually, in this example, the test with the larger standard deviation would contribute almost twice as much to the final score determination as the test with the smaller standard deviation. You can easily control for differences in dispersion across tests by converting to standard scores. You are, in effect, equalizing the standard deviations and means of both tests. With this accomplished, you can add the standard scores of both tests together and assign final grades.

Association

Determining the extent of association is the goal of many statistical analyses. Correlational statistics allow us to state quantitatively the degree of association between two measured variables. These variables can be two sets of test scores, or more commonly, a predictor test and a criterion score. The higher the correlation between the two variables, the greater the relationship or association between them. That is, we can with greater certainty predict the scores on variable Y with knowledge of the scores on variable X. Correlation coefficients are usually depicted as decimal numbers which vary between + 1.00 and - 1.00. A correlation of + 1.00 is a perfect positive correlation, while a correlation of - 1.00 is a perfect inverse correlation. A correlation coefficient of 0.00 indicates no relationship between the two variables under consideration. Both positive and inverse coefficients can be useful. Actually, it is the size, not the sign, which is of importance. As both positive and inverse correlations approach zero, then they become less useful for predictive purposes.

Whether or not a given correlation coefficient is significant beyond chance depends on the number of cases on which it was based. As the number of cases increases, the magnitude of the correlation you need for statistical significance decreases.

There are several types of correlation coefficient. Which type is used depends on the type of data you have collected. For instance, if you wish to obtain a Pearson Product Moment correlation coefficient, your score distributions must be continuous. If your data is in the form of ranks, the Spearman Rank Order correlation coefficient is the most appropriate. If one distribution is continuous and one is dichotomous, you can use either the Biserial or the Point-Biserial coefficient. Finally, if the data for both of your variables are in the form of categories, you should use either the Phi or the Tetrachoric coefficient.

These latter correlation coefficients are statistical approximations of the Pearson Product Moment coefficient (r), which is the most standard one. Calculation of a Pearson r is shown in the next section. There will be many instances, though, when you will not be able to use Pearson r, and under these circumstances, you will have to choose one of the other coefficients. For purposes of illustration, the calculation of two of these coefficients, the rank order and the Phi coefficients, will be considered in detail.

Pearson Product Moment Correlation Coefficient

The calculation of the Pearson Product Moment Correlation Coifficient or Pearson r can be readily accomplished on a calculator. A step by step calculation of Pearson r is shown in Exhibit 29.

Exhibit 29

Example of the Calculation of a Pearson Product Moment Correlation Coefficient

Raw Scores Variable X	Raw Scores Variable Y	X^2	Y^2	XY
7	6	49	36	42
11	8	121	64	88
9	8	81	64	72
5	4	25	16	20
8	7	64	49	56
9	5	81	25	45
10	9	100	81	90
7	5	49	25	35
6	6	36	36	36
12	11	144	121	132
Σ 84	69	750	517	616

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{N\Sigma X^2 - (\Sigma X)^2}\sqrt{N\Sigma Y^2 - (\Sigma Y)^2}}$$

$$= \frac{10(616) - (84)(69)}{\sqrt{10(750) - (84)^2}\sqrt{10(517) - (}} =$$

$$= \frac{6160 - 5796}{\sqrt{7500 - 7056}\ \sqrt{5170 - 4761}}$$

$$= \frac{364}{\sqrt{444}\ \sqrt{409}}$$

$$= \frac{364}{(21.1)(20.2)}$$

$$= \frac{364}{426.2}$$

$$= .85$$

Rank Order Correlation

You can use the Spearman Rank Order coefficient (Spearman Rho, ρ), if your data are in the form of ranks, or if you want a quick approximation to the Pearson r. In this latter case, you will have to convert your raw data to ranks. Three short examples are presented in Exhibits 30, 31, and 32. One shows a strong positive relationship (Exhibit 30), the second shows a strong inverse relationship (Exhibit 31), and the third example shows no relationship between the variables under consideration (Exhibit 32).

The examples also illustrate the procedure to follow if ties exist in the data. If the two highest scoring persons have the same score, you simply take the average of the first two ranks [(1 + 2)/2 = 1.5] and assign this value to each of them. If the two persons after the first two are tied, each of them receives a rank of 3.5 [(3 + 4)/2 = 3.5]. If the first three persons are tied, all would receive a rank of 2 [(1 + 2 + 3) /3 = 2.0].

Finally, ρ should not be used if the number of subjects in your sample is greater than 30. Ranking can be quite tedious under these circumstances. The only exception to this rule of thumb is when your data are already in rank form.

Exhibit 30

Example of Spearman Rank Difference Correlation with a Strong Positive Relationship between Variable X and Variable Y

Raw Scores Variable X	Raw Scores Variable Y	Rank X	Rank Y	Difference (d)	d^2
7	6	7.5	7	.5	.25
11	11	2	1	1.0	1.00
9	8	4.5	4	.5	.25
5	4	10	10	.0	.00
8	7	6	5.5	.5	.25
9	7	4.5	5.5	1.0	1.00
10	9	3	3	.0	.00
7	5	7.5	8.5	1.0	1.00
6	5	9	3.5	.5	.25
12	10	1	2	1.0	1.00

$5.00 = \Sigma d^2$

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(5)}{10(100-1)}$$

$$= 1 - \frac{30}{990}$$

$$= 1 - .03$$

$$= .97$$

Exhibit 31

Example of Spearman Rank Difference Correlation with a Strong Inverse Relationship between Variable X and Variable Y

Raw Scores Variable X	Raw Scores Variable Y	Rank X	Rank Y	Difference (d)	d^2
7	7	7.5	4.5	3.0	9.00
11	4	2	9	7.0	49.00
9	6	4.5	6.5	2.0	4.00
5	12	10	1	9.0	81.00
8	7	6	4.5	1.5	2.25
9	8	4.5	2.5	2.0	4.00
10	3	3	10	7.0	49.00
7	6	7.5	6.5	1.0	1.00
6	8	9	2.5	6.5	42.25
12	5	1	8	7.0	49.00

$290.50 = \Sigma d^2$

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(290)}{10(100-1)}$$

$$= 1 - \frac{1740}{990}$$

$$= 1 - 1.76$$

$$= -.76$$

Exhibit 32

Example of Spearman Rank Difference Correlation with No Relationship between Variables X and Y

Raw Scores	Raw Scores	Rank	Rank	Difference	
Variable X	Variable Y	X	Y	(d)	d^2
7	4	7.5	8	.5	.25
11	8	2	4	2.0	4.00
9	5	4.5	6.5	2.0	4.00
5	9	10	2.5	7.5	56.25
8	3	6	9	3.0	9.00
9	10	4.5	1	3.5	12.25
10	2	3	10	7.0	49.00
7	6	7.5	5	2.5	6.25
6	5	9	6.5	2.5	6.25
12	9	1	2.5	1.5	2.25

$149.50 = \Sigma d^2$

$$\rho = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

$$= 1 - \frac{6(149.50)}{10(100-1)}$$

$$= 1 - \frac{897}{990}$$

$$= 1 - .91$$

$$= .09$$

The Phi Coefficient

The Phi coefficient can be derived from the statistic known as Chi^2 (χ^2). The Phi coefficient is used when your data are dichotomous (e.g., pass-fail) on each of the two variables under consideration. To calculate χ^2 and Phi, you first construct a 2 x 2 table for your data. Assume that you want to determine if performance in ground training is associated with flight training performance. To do this, you have to obtain data relative to each of the following four categories:

1. the number of persons who performed adequately in both ground training and flight training
2. the number of persons who performed adequately in ground training and poorly in flight training
3. the number of persons who performed inadequately in ground training and poorly in flight training
4. the number of persons who performed inadequately in ground training and adequately in flight training.

Your 2 x 2 table would look something like the one pictured below. Each numerical entry represents the number of persons in each category.

		Ground Training	
		Adequate	Inadequate
Flight Training	Adequate	63	25
	Inadequate	12	47

Inspection of the table indicates that some association exists. That is, knowing that a person's ground training performance is adequate allows you to predict with a fair degree of accuracy that his flight training performance will be adequate. But what is the degree of association between the two variables?

The next step is to calculate the expected value in each cell. To do this you sum across all rows and columns, as below:

		Ground Training		
		Adequate	Inadequate	
Flight Training	Adequate	63	25	88
	Inadequate	12	47	59
		75	72	147

The expected values are then calculated as follows:

adequate : adequate	(88 x 75)/147 = 44.9
adequate : inadequate	(88 x 72)/147 = 43.1
inadequate: adequate	(59 x 75)/147 = 30.1
inadequate: inadequate	(59 x 72)/147 = 28.9

Your table will now look like the following, where the expected values are entered in parentheses:

		Ground Training	
		Adequate	Inadequate
Flight Training	Adequate	63 (44.9)	25 (43.1)
	Inadequate	12 (30.1)	47 (28.9)

The χ^2 value is obtained by taking the frequency observed (f_0) minus the frequency expected (f_e) minus one half, squaring this value, and dividing by the frequency expected for each cell. This operation is shown below:

$$\chi^2 = \frac{[(f_o - f_e) - .5]^2}{f_e}$$

$$= \frac{[(63-44.9)-.5]^2}{44.9} + \frac{[(25-43.1)-.5]^2}{43.1} + \frac{[(12-30.1)-.5]^2}{30.1}$$

$$+ \frac{[(47-28.9)-.5]^2}{28.9}$$

$$= \frac{[17.6]^2}{44.9} + \frac{[17.6]^2}{43.1} + \frac{[17.6]^2}{30.1} + \frac{[17.6]^2}{28.9}$$

$$= \frac{309.76}{44.9} + \frac{309.76}{43.1} + \frac{309.76}{30.1} + \frac{309.76}{28.9}$$

$$= 6.91 + 7.19 + 10.29 + 10.72$$

$$= 35.11$$

The above value is significant enough to occur by chance only one time in one thousand. The formula for determining the phi (ϕ) coefficient is:

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

In our example this becomes:

$$\phi = \sqrt{\frac{35.11}{147}}$$

$$= \sqrt{.238}$$

$$= .49$$

The final correlation is, of course, .49. The above procedure seems involved and tedious, but if you look closely at the steps, you will notice many shortcuts. The procedure can also be extended to tables larger than 2 x 2 with the restriction* that there be at least five or preferably ten subjects per cell. Another difference with tables larger than 2 x 2 is that you don't have to subtract .5 from the numerator in each cell. In this case, your formula becomes:

*This restriction also applies to the 2 x 2 table.

$$\chi^2 = \frac{(f_o - f_e)^2}{f_e}$$

The correlation coefficient derived from tables using the above formula is called the contingency coefficient. The formula for the contingency (C) coefficient is:

$$C = \frac{\chi^2}{n + \chi^2}$$

Degree of Freedom

Any understanding of statistical methods depends on an understanding of the "degrees of freedom" concept. Prior to determining if your χ^2 value or contingency coefficient is statistically significant, you must determine its associated degrees of freedom (d.f.). The degrees of freedom is a function of the number of rows and columns in your χ^2 table. The formula for degrees of freedom is as follows:

$$\text{d.f.} = (\text{rows} - 1)(\text{columns} - 1)$$

If you had a 3 x 3 table, your degrees of freedom would be 4, and with a 4 x 5 table, the degrees of freedom are 12. You must know the degrees of freedom because the χ^2 value needed for statistical significance decreases as the degrees of freedom increase.

Any time you wish to know the level of confidence that you can place in a statistic, you must know the number of degrees of freedom involved. There is a different method for determining the number of degrees of freedom for each statistic. Accordingly, you should consult a statistical text for advice on the number of degrees of freedom you have when you calculate other statistics.

Inferential Statistics

The main difference between correlational and inferential statistics is that the latter allow you to determine cause, while the former only allow the determination of association. In correlational statistics, you measure and correlate what already exists in the student. This can include but is not limited to: (1) written test scores, (2) ratings, (3) performance test scores, and (4) background data. All you can determine is whether or not one variable covaries with another. Correlation in no way implies causality. This is not to say that there is not cause, but that you cannot determine cause from a correlational analysis. Inferential statistics, on the other hand, impose experimental conditions upon the subjects and allow you to determine cause. For example, you may have two matched samples of students. One sample may be trained using method A while the other sample is trained using method B. You may find, after training, that students taught by method B perform electronic trouble shooting better than students taught by method A. You can conclude from these results, all other factors being equal, that method B is better than method A for teaching electronic trouble shooting.

The essential difference, then, between correlational and inferential statistics is that the latter imposes experimental manipulations upon subjects, while the former just measures the subjects. You may feel that other variables, in addition to the teaching method, are important in the learning situation.

Independent and Dependent Variables

One variable (the variable you measure) will always be called the dependent variable, while the others are all called independent variables. In this example, the two methods of training are the independent variables and student trouble shooting time is the dependent (response) variable. You can have only one dependent variable, but any number of independent variables. Generally, when you have more than one independent variable, you will need a statistician or a psychologist to help accomplish the statistical analysis.

Interaction

Consider an experiment which investigates the effects of varying class size and method of instruction (independent variables) on trouble shooting time (dependent variable). Your data might look like this.

	Large Size Class	Medium Size Class	Small Size Class	
Method A	$\bar{x} = 20$	$\bar{x} = 28$	$\bar{x} = 30$	26.0
Method B	$\bar{x} = 19$	$\bar{x} = 35$	$\bar{x} = 37$	29.7
	19.5	31.5	33.5	

The entries in each of the cells in this table consist of mean posttraining trouble shooting performance scores of the trainees. If you inspect the row and column averages, you might conclude that method B is better than method A and that small and medium size classes are better than large classes. This conclusion is incorrect. If you look at the individual cell entries, you will notice that there is an interaction between teaching methods and class size, such that in large class sizes teaching method is inconsequential. Only in medium and small size classes does teaching method exert a differential influence. The interaction concept is a singularly important one to understand. Generally, an interaction exists when one independent variable (e.g., class size) has a differential effect across various levels of another independent variable (e.g., teaching method). That is, the effects are not additive. If the mean in the method A, large size class cell were 14 rather than 20, an interaction would not have existed. The effects of the two independent variables would have been additive, and you could use the column and row means to draw conclusions.*

In order to determine if an interaction exists in your data, it is always best to plot the means on a graph. Figure 2 presents just such a plot for our hypothetical data.

*After an appropriate statistical analysis.

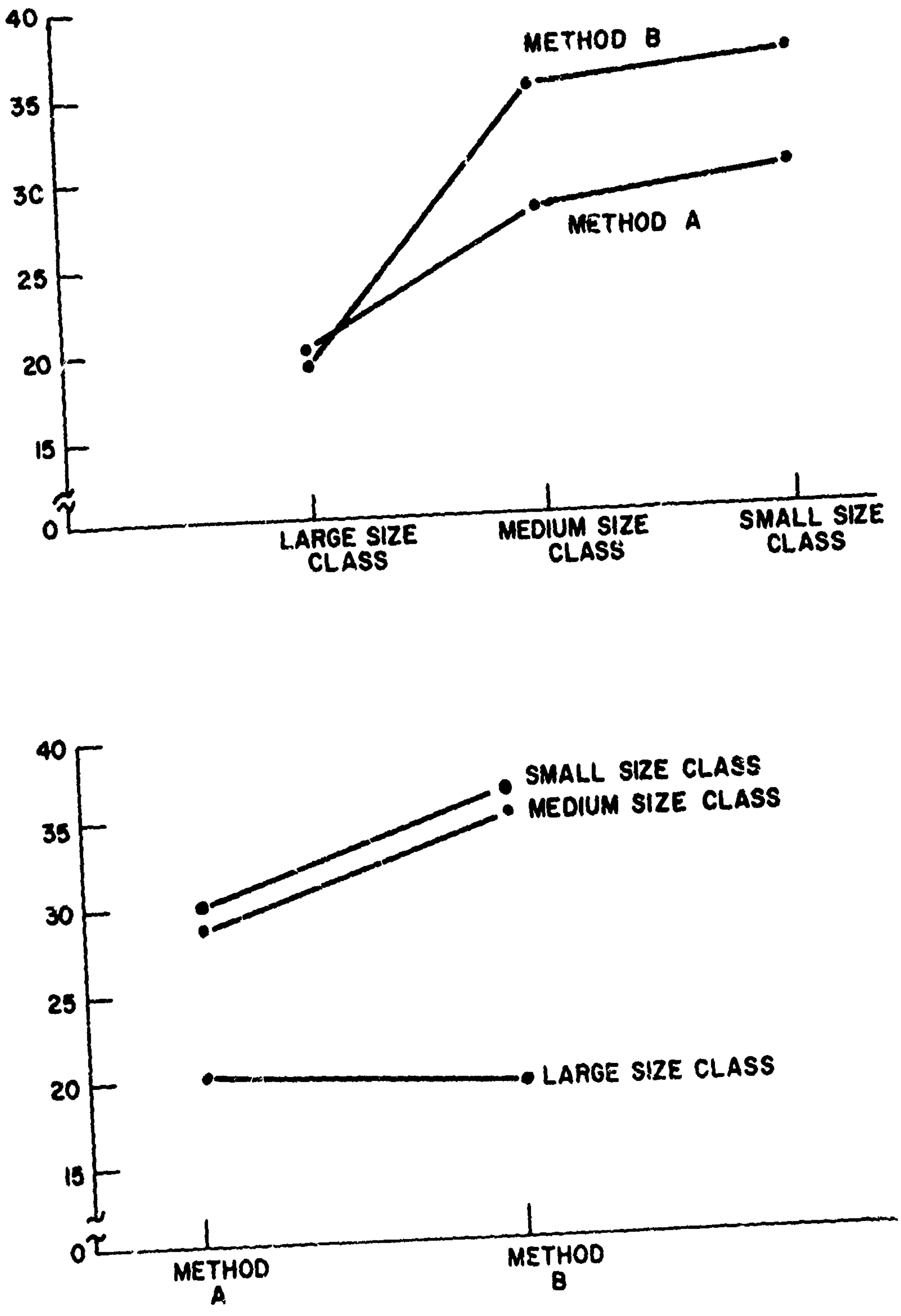


Figure 2. Plot of interactions for hypothetical data.

Plots like these can always tell the investigator a great deal about the results of his experiment. If the lines deviate from parallel by an appreciable amount, you can be fairly certain that an interaction exists in your data. In a sense, plotting a graph tells you what to expect from your subsequent statistical analysis.

Analysis of Variance

When an experiment is performed which contains several independent variables, the statistical analyses which is performed is called an analysis of variance (ANOVA). If an experimental manipulation is involved, ANOVA will help you determine cause. If you do not impose experimental conditions, though, ANOVA will not allow you to interpret your results in terms of causation. For example, if your independent variable is a test score you will be determining association, not causation. On many occasions you may want to mix experimental variables with non-experimental or classification variables, because you feel that the classification variable (e.g., age, education, sex, race, test score, etc.) will have an effect on the dependent variable. Some might argue that these classification variables can be considered experimental manipulations in the sense that society or nature imposes them. This argument is moot since so many factors can enter into the resultant effects of the classification variable as to make the determination of causation untenable. For instance, race might be associated with poorer posttraining performance, but this does not mean that race caused the poorer performance. Actually, cultural deprivation may be a factor common to both race and posttraining performance, and this maybe the real causitive agent. The methods for performing an analysis of variance are beyond the scope of this manual but are available in many statistical texts. Generally, this method of analyzing data is one of the most powerful available and studies should be designed with the use of this method in mind.

Statistical Control

There are several meanings for the term control as it is used in evaluation, experimentation, and statistics. We shall concern ourselves, though, with only two of the more important types, statistical control and experimental control. Statistical control attempts by arithmetic methods, to remove from the data certain effects which may contaminate your results. Experimental control attempts to achieve the same purpose, through manipulation of the research situation. This section considers statistical control; experimental controls is discussed in the section which follows.

You may have two variables (e.g., height and weight) which have a higher correlation than expected due to the effects of a third variable (e.g., age). If the effects of age were removed, then the correlation between height and weight would be reduced. Such a statistical control procedure exists; it is called partial correlation. An analagous statistical control procedure is known as analysis of covariance (ANCOV). This is simply a specialized type of ANOVA procedure. For example, you might have two classes, one taught by method A and one taught by method B. As in our previous example, you might find that method B yields superior posttraining performance. It may also be the case that the average intelligence of the subjects taught under method B was considerably higher than the average intelligence of the subjects taught under method A. You will then want to use a measure of intelligence as a covariate in order to adjust the posttraining performance scores for intelligence. If the proper adjustment and analysis are performed, method A may, indeed, be equivalent to method B.

The above procedures are known as statistical controls. A number of effects can not easily be removed from the data through statistical methods. To control for these effects, the research situation itself must be manipulated.

Other types of factors may also act to confound your data and should be controlled experimentally. Examples are:

1. History or Antecedents

The past history or antecedent conditions of all persons assigned to different experimental groups or treatments should be roughly comparable. If this precaution is not exercised, the results of the experiment are apt to be meaningless.

2. Maturation and Motivation of Subjects

The level of maturity of all the subjects in an experiment should also be controlled. Skilled behaviors can change in subjects from infancy through old age. If maturity level is confounded with experimental measurement, the results can be misleading or worthless. Similarly, the interest level or the motivation of those who are used as subjects can confound the results of your study. The motivation of those assigned to different experimental groups should be equal and all persons should be given the same explanation regarding the purposes of an experiment.

3. Testing Effects

Some of the subjects in your experimental groups may have had differential prior test taking experiences. Less sophisticated persons who are unfamiliar with tests and test taking are likely to be overanxious and attain test scores which are underestimates of their true learning or ability. Those individuals who have had little testing experience should be given special consideration, or even practice in test taking; the Psychological Corporation publishes a tape and booklet series entitled How to Take Tests which is acceptable for this purpose.

In a similar vein, the very act of testing can alter or interfere with the experimental variables under consideration. This is the well known "Hawthorne" effect. One way to avoid this problem is to refrain from telling the subjects that they are a part of a research study until after the data have been collected.

4. Instrumentation

Sometimes the use of instruments or gadgetry in the experimental situation can arouse anxiety or fear in the subjects. If instrumentation is employed, the doubts, fears, and anxieties of the subjects should be allayed before any experimentation begins.

Finally, the instrument or apparatus used should provide usable data. Some instruments (e.g., polygraphs) yield information which is very difficult to interpret. Such instruments should be avoided, unless you are an expert in their use.

5. Pretest Sensitization

Sometimes, as an experimenter, you will have the opportunity to use a pretest - posttest experimental design. Pretesting a subject can often sensitize him so that his post experimental test score is affected. This effect is often the reason for avoiding experimental designs which are based on a pretest-posttest paradigm.

6. Varying Environmental Conditions

All aspects of the environment, except the independent variable which is systematically manipulated, must be kept constant for each of the experimental groups. Otherwise it will not be possible to know whether or not any differences noted are due to the manipulation of the independent variable or to the environmental conditions which vary across the groups.

7. Effects of Prior Treatments

Persons who have served as subjects in one or more previous experiments are often so sophisticated that they can bias your experimental results. Often, they will try to guess your intent and give you the results they think you want, or vice versa. Others, will be overly suspicious and suspect deception at every turn. This type of person is a poor risk in any experimental context. If you cannot obtain "naive" or inexperienced subjects for your experiments, you might try to assess suspiciousness and experimental sophistication through post-experiment questionnaires and interviews.

Another direct control technique is through repeated measures. Repeated measures involve measurement of all subjects across all levels of the independent variable(s) in question. For example, you might be interested in measuring the performance of skilled craftsmen on two new pieces of equipment. You might have all craftsman perform on both pieces of equipment, rather than splitting the sample in half with each half using only one piece of equipment. In this way you have controlled for possible chance variations between samples. In such an experiment you would need further controls to account for such contaminants as sequential effects and practice. One way to do this is to have half of the subjects first perform on equipment A while the other half starts on equipment B. Another way is to employ an ABBA order. The traditional learning experiment in which a student's performance is followed over several learning trials is a variation of the repeated measures paradigm.

We can conclude that experimental and direct control procedures are superior to statistical control procedures. The latter should only be used if direct control can not be imposed on the study.

Replication

If, in your experiment, you obtain some important results, you should subject these results to verification. The best way to do this is to repeat (replicate) your experiment on a different sample of subjects. If replication is not possible, you should, at least, describe your experiment in a way that will allow others to duplicate your experimental conditions.

More Advanced Statistical Methods

Regression

The technique of constructing a regression equation is beyond the scope of this manual. Suffice to say that the regression equation utilizes the predicter-criterion correlation in order to predict future behavior. For example, the correlation between predictor test scores and school scores can be used to predict school performance (criterion). This is to be done, of course, only after the original correlation has been validated a second time.

There is also a technique known as multiple regression in which several tests can be combined in order to predict a single criterion. Inclusion of more than four or five predictors in the equation is usually wasted effort, because the resultant multiple correlation coefficient will not be affected to any great extent by the additional predictors. Your first four tests might give you a multiple correlation of .50, and the addition of four more might only raise it to .52 or .53. It is hardly worth the extra time and effort to include the additional four variables. The reason the added variables do not account for much predictable behavior is probably because they overlap with the first four to such an extent that they contribute no additional unique prediction of their own.

Expectancy Tables

To aid in the interpretation of a correlation coefficient, it is recommended that you use expectancy tables. An expectancy table is a special graphic or tabular presentation of the relationship between the predictor and the criterion.

There are several methods for constructing an expectancy table. One method is to count the proportion of people who achieve each predictor score and who are above the 50th percentile on the criterion score. Another, and perhaps better way, is to tabulate the number of persons who perform satisfactorily at each predictor score.

The first step in constructing an expectancy table is illustrated in Exhibit 33.

Exhibit 33

Example of the First Step in Constructing an Expectancy Table

Predictor Test	Criterion Performance	
Score	Successful	Unsuccessful
20	2	0
19	1	0
18	4	1
17	2	2
16	4	3
15	8	3
14	5	4
13	5	5
12	4	5
11	2	4
10	2	3
9	2	2
8	2	2
7	0	2
6	0	2
5	0	2

After counting the number of persons achieving each predictor score who were either successful or unsuccessful on the criterion performance, the proportion of persons who were successful or unsuccessful in score ranges is tabulated. For example, for people in the 17-20 score range, 75 per cent were successful. In the prediction situation, you would say that a man whose score fell in this range has 75 chances in 100 of being successful. A completed expectancy table, based on the data of Exhibit 33, is presented as Exhibit 34.

Exhibit 34

Example of a Final Expectancy Table

Predictor Test Score	Probability of Successful Criterion Performance*
17-20	75
13-16	59
9-12	42
5 - 8	20

* Chances in one hundred.

The cutoff score on the predictor test will depend on Air Force Manpower needs. But, for our example, only under very unusual circumstances would it be below the 9 to 12 score range.

If you have a number of predictor tests, the multiple cutoff or successive hurdles approach may be used for predictive purposes. In the multiple cutoff method, you set a realistic minimal score for each test which all applicants or trainees must exceed. In setting cut scores, you should remember that the scores should be realistically related to job performance.

You may decide to allow some leeway in this method. If you administer four tests, you can make the stipulation that the applicant or trainee pass the cut score on only three.

In the successive hurdles method, the trainee or applicant must pass each test in sequential order. If he fails at any one step or level, he is not permitted to go on to the next level. Usually, each step or level is arranged in order of importance. Therefore, some ranking or weighting of the tests is required. Successive hurdles is a variation of the multiple cutoff method, and the same precaution regarding the setting of cut scores should be observed.

Comparison of the Multiple Regression and Multiple Cutoff Methods

In this report, we have suggested that you use one of the variations of the multiple cutoff method rather than the multiple regression method. The main reason is that the multiple cutoff method only requires the calculation of a measure of association (e.g., product moment correlation), while the multiple regression approach requires some very complex calculations. Beyond these reasons, the choice of method depends on your own philosophic preference and the demands of the job or training situation. You should understand the basic differences between these methods so that you can choose more intelligently between them. In multiple regression, those variables which account for a greater proportion of the criterion behavior are weighted more heavily in the prediction equation. Predictor variables which are relatively more independent from the other predictor variables are also weighted more heavily in the equations. Also, in multiple regression, a high score on one test can nullify a low score on another. Such is not the case in the multiple cutoff method, unless you allow the trainee to fail one of the tests. This is the basic difference between the methods. If you feel that a man should not fail some of your tests, then you should use the multiple cutoff method. On the other hand, if you think that a high score on one test can compensate for a poor score on another test, then you can use multiple regression.

Curvilinear Relationships

One procedure you should always perform, prior to calculating a correlation coefficient, is to plot your data on a graph. The reason for graphing the data is to determine if there is any curvilinear relationship present. This is a necessary procedure, since most correlation coefficients (all those we have discussed) are only sensitive to linear relationships in the data. One common occurrence is the case in which persons who score in the middle range of a predictor perform best on the job, and those who score at the extremes perform poorly. In this case, the product moment correlation, or any other correlation coefficient which yields linear predictability is useless. A pictorial presentation of this situation is presented in Figure 3.

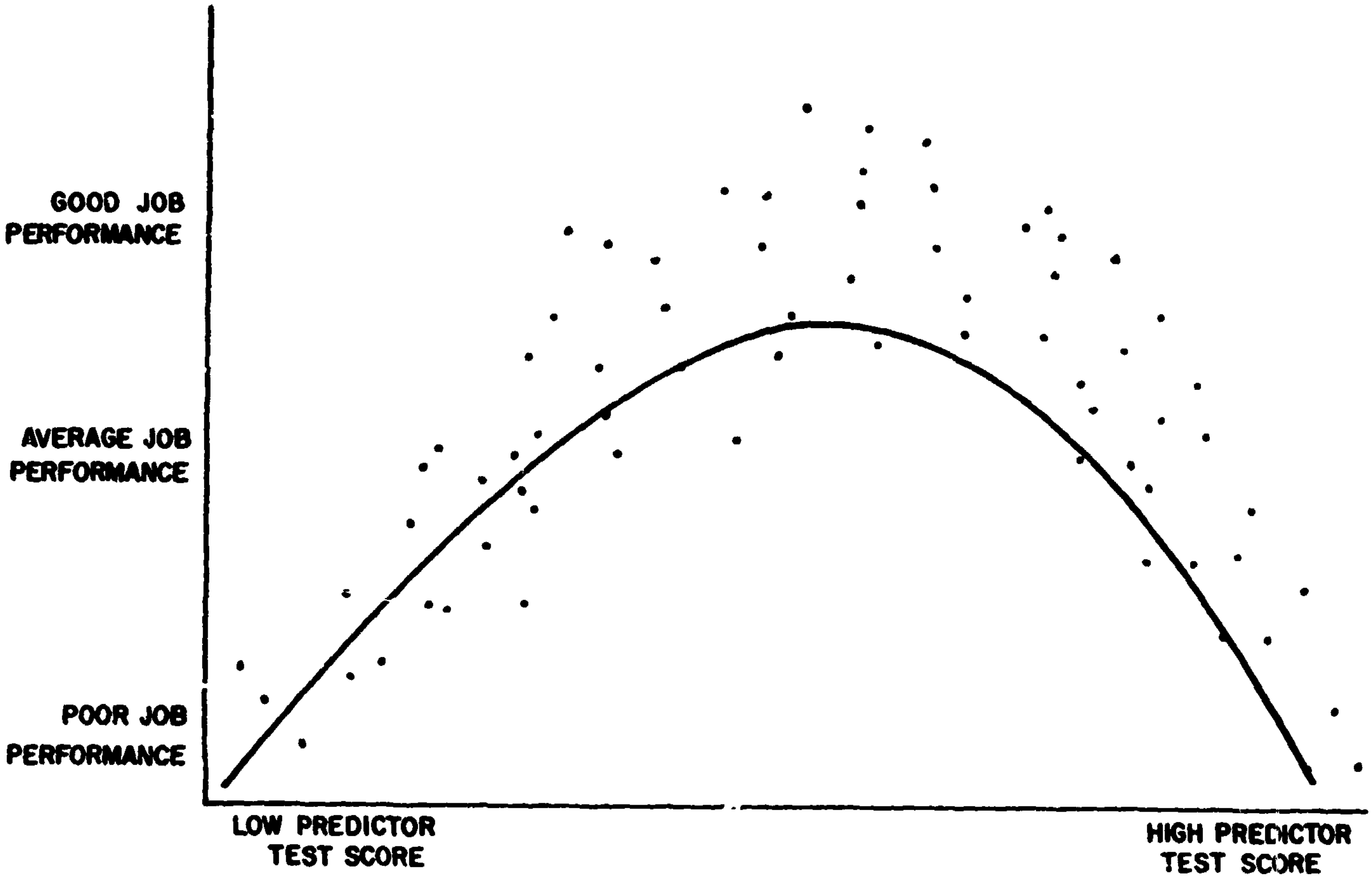


Figure 3. Example of curvilinear relationship.

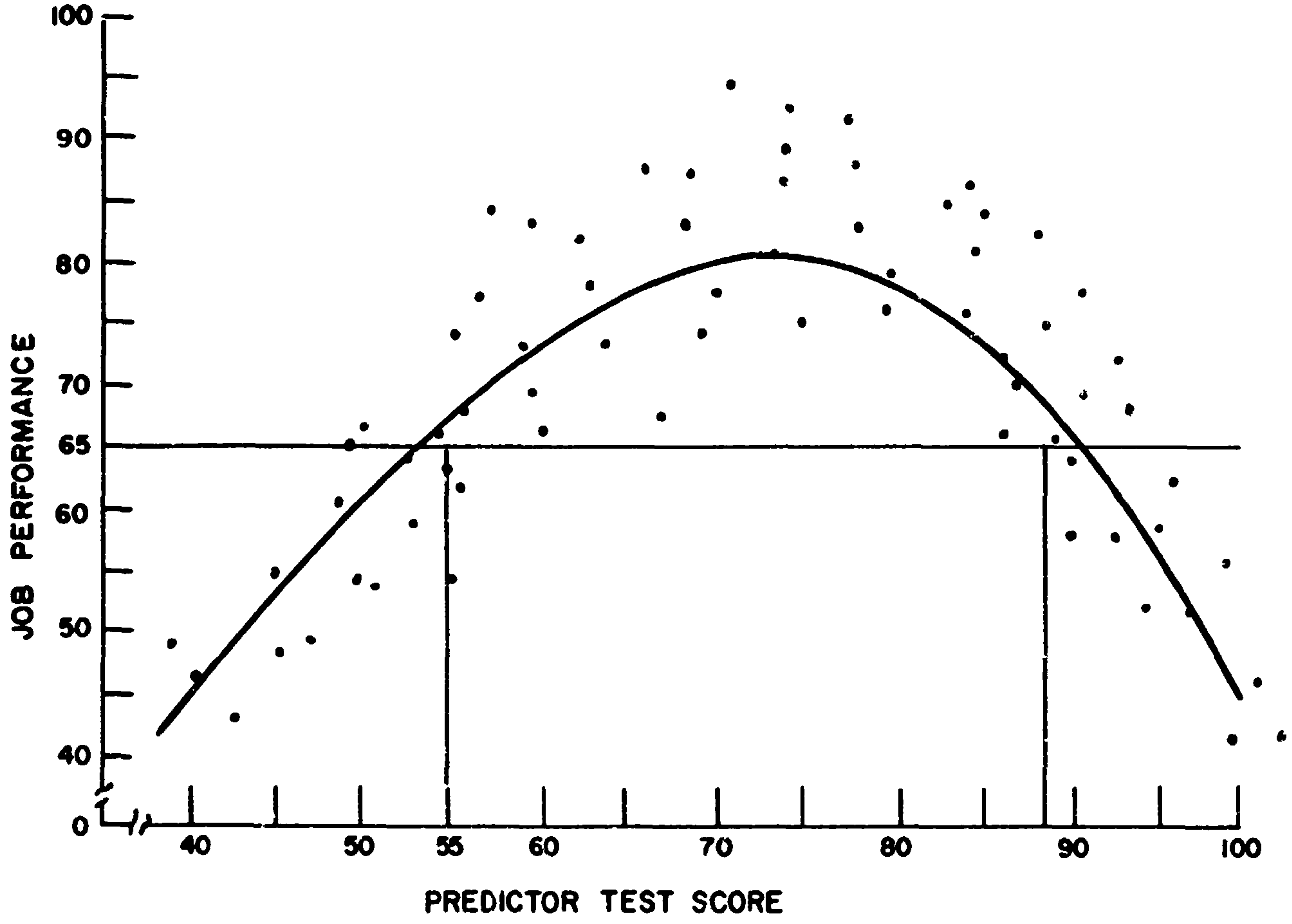


Figure 4. Example of application of manual fit cut scores to curvilinear relationship.

A correlation coefficient is needed in this situation which gives you the best fitting line rather than the best fitting linear relationship. The eta coefficient can be calculated for this purpose. However, calculation of eta can be avoided if your graphic plot definitely shows a curvilinear relationship. For example, let us assign numbers to the points in the previous illustration, as in Figure 4. After you plot the data, you determine that point on the job performance dimension representing minimally adequate job behavior. After careful analysis, you might decide that a job performance score of 65 is minimally adequate. Then draw a line horizontally across the graph paper at a performance score of 65. Where this line intersects the data, draw a vertical line down to the predictor test score axis. In the example, persons whose predictor test scores fell in the 56-88 range are most likely to be successful. Persons who scored above 88 (cut score 1) and below 56 (cut score 2) are less likely to be successful. This kind of situation often exists in clerical jobs which require only a moderate amount of ability. Persons with low test scores perform poorly, because they lack the requisite clerical ability, while persons with high test scores perform poorly because they are bored with the routine nature of the work. Curvilinear relationships can exist in other kinds of associations as well.

Averaging Coefficients of Correlation

Correlation coefficients should not be averaged or otherwise treated arithmetically. Coefficients of correlation are index numbers. They are not equal values along a scale. For example, a predictor-criterion relationship of .80 does not mean that this relationship is twice as strong as a relationship of .40. A correlation of .80 accounts for four times as much variance as a correlation of .40. Before treating correlation coefficients arithmetically (e.g., averaging them), they should first be converted to standard scores. There are special tables for this conversion in most introductory statistics text books. You simply look up the standard score (z value) corresponding to each coefficient to be included in your calculations. Perform your arithmetic operations; then, through use of the same table, transform the final z value back to a correlation coefficient.

Often, you will want to average correlation coefficients from subsamples to obtain an overall estimate of a population correlation. An example of such a calculation is presented in Exhibit 35.

Exhibit 35

Example of a Calculation Involving Averaged Correlation Coefficients

Coefficients to be Averaged	z Value
.88	1.38
.27	.28
.72	.91
.51	.56
.37	.39
.92	1.59
.20	.20
.46	.50
.63	.74
.14	.14
Sum of z values	6.69
Average z value	.67
Average correlation	.58 (from table)

Factor Analysis

Factor analysis is simply a statistical method for eliminating the redundancy present in a correlation matrix. A correlation matrix is a row by column arrangement of all possible correlations among a group of predictors in a set. One might, for example, be able to reduce a 20 x 20 correlation matrix to a 20 x 5 factor matrix, thus using only five factors rather than 20 items to describe the matrix. Factor analysis is essentially a grouping procedure which quantitatively brings together or clusters correlated groups and differentiates them from other groups. For example, if you performed a large group of measurements on a box, intercorrelated the measurements, and factor analyzed the correlation matrix, you would probably isolate three dimensions: length, width, and height. The factor analytic procedure is most useful where the dimensionality of the area of interest is unknown. The technique has been employed, for example, to order such domains as personality and intelligence; an example of a dimension derived from a factor analysis of personality is "introversion." Each dimension or factor which is isolated in a factor analysis is consistent and orthogonal (uncorrelated) to the other dimensions or factors derived from the matrix. The factors are rendered orthogonal (independent) by a process called rotation. Generally, items which are highly correlated will fall in the same factor, while items that are uncorrelated with each other will appear in different factors. Factor analysis is a complex procedure which can take hundreds of hours if performed manually. Fortunately, high speed computers can perform a factor analysis in minutes. Thus, the technique is available to almost anyone performing research.

In order to acquaint the reader with the procedure, a sample correlation matrix (Exhibit 36) and factor matrix (Exhibit 37) is presented for a 10 item attitude scale. These are actual results obtained from a research project using 144 student subjects.

Exhibit 36

Item by Item Correlation Matrix of a 10 Item Attitude Scale

Item Number	Item Number								
	2	3	4	5	6	7	8	9	10
1	.48	.41	.38	.39	.43	.45	.30	.25	.53
2		.43	.42	.48	.25	.31	.23	.24	.41
3			.51	.39	.29	.26	.43	.22	.27
4				.34	.44	.40	.39	.45	.29
5					.32	.30	.53	.17	.63
6						.52	.27	.35	.35
7							.36	.38	.36
8								.32	.51
9									.22

Exhibit 37

Factor Matrix and Cumulative Proportions of the Total Variance of a 10 Item Attitude Scale

Item	Factor Loadings			
	1	2	3	4
1	.34	.30	.18	.54
2	.19	.16	.34	.51
3	.17	.26	.59	.25
4	.47	.17	.52	.20
5	.15	.68	.20	.19
6	.58	.20	.12	.23
7	.60	.22	.10	.24
8	.27	.59	.31	.01
9	.52	.11	.23	.05
10	.21	.68	.04	.39
Cumulative Proportion of Total Variance	15.06	30.90	40.60	50.13

Essentially, the results in Exhibit 37 tell us that four dimensions can adequately describe the correlation matrix, rather than 10 dimensions or items. A factor loading is the correlation coefficient of an item with a factor or dimension. For example, item two is correlated .19, .16, .34, and .51 with each of the four factors, respectively. By convention, factor loadings between .3 to .4 or higher are considered meaningful. You can see from Exhibit 37 that items 4, 6, 7, and 9 load the highest on factor 1. All of these items were concerned with hostility, lack of cooperation, and unfriendliness. From this, the factor was named "hostility". Generally, the factor analyst names each factor in accordance with the item constellation that composes it. You will also notice, at the bottom of Exhibit 37, a row listing the cumulative proportions of the total variance. This represents the extent that the factors account for the total variation in the original correlation matrix. The amount of variance accounted for by each factor is calculated by squaring and summing the factor loading for each factor. For factor 1 this figure was 15.06. The amount of variance accounted for by each factor is also a convenient method for deciding if a factor is large enough to include in your results. In some factor analytic computer programs you can specify, in advance, how many factors you want calculated. In others, factors are continuously extracted until all the predictable variance is accounted for. In this case, there are a number of decision rules to follow for deciding how many of the factors represent an adequately comprehensive and economical description of the data set. In the example cited, the fifth and sixth factors only accounted for two or three per cent of the variance, while each of the first four factors accounted for 10 to 16 per cent of the variance. On this basis, it was decided that four factors adequately describe the correlation matrix. One good method for deciding where to stop factoring, is to graph the amount of variance accounted for by each factor and note when there is a sudden drop off. There are other more sophisticated procedures which can help you to decide when to stop factoring, but they are beyond the scope of this handbook. If at all possible, you should consult a quantitatively oriented psychologist before making any final conclusions about your factor analytic results.

Obviously, factor analysis can be a highly useful tool in training evaluation and student achievement measurement. For example, you might have a 15 item rating scale which measures the on-the-job behavior of training school graduates. It would be unwieldly to describe the on-the-job behavior of these men in terms of either the 15 separate dimensions or one overall composite, when the 15 item scale may possibly be reduced to three or four dimensions which describe on-the-job behavior. If predictor tests were used, then, significant validity coefficients might be dependent upon whether or not one used factor analysis.

Q Factor Analysis

Another old technique, but one which will probably be used more frequently during the next decade, is Q factor analysis. In performing a Q factor analysis, one simply factor analyzes the matrix of person correlations rather than item correlations. This method can be useful for grouping persons who think or behave similarly. For example, when constructing a training program, it may be useful to know the different cognitive styles of the potential trainees so that the training could be adapted to the needs of each homogeneous group.

Canonical Correlation

Canonical correlation is an extension of factor analysis to the situation in which two separate sets of variables exist. The first canonical correlation is the highest correlation between a factor of the first set of variables with a factor from the second set of variables. The second canonical correlation is the correlation between a second factor of the first set of variables with a second factor of the second set of variables. Canonical correlations are continually extracted until all the common variance between both sets of variables is accounted for. The method is most applicable when there are two separate sets of variables, for example, one set of predictor variables and one set of criterion variables. Here, instead of correlating predictor tests with criterion scores, canonical correlation allows one to correlate predictor factors with criterion factors.

Moderator Variables

A test is a moderator variable when its scores differentially determine the predictability of another test or variable. For example, one may be able to predict very adequately the performance of college students using an intelligence test for those who score high on a test of achievement motivation, but not for those who score low on the test of achievement motivation. Race is one of the more currently popular moderator variables. Much recent research has shown that employment tests are differentially predictive across racial groups. This finding supports the contention that common selection standards for negroes and whites are inappropriate or unfair. Examples of other factors which can be used as moderators are: (1) achievement level, (2) personal and environmental variables, (3) social background factors, (4) cognitive styles, and (5) emotional reactions.

It is easy to determine if a variable is a moderator. First, you separate those who score high on the moderator from those who score low on the moderator. Splitting the group in half (or thirds) will suffice. Then observe whether your remaining variables differentially predict performance across the high and low groups. If a math test predicts future performance for those who score high on achievement motivation, but not for those who score low on achievement motivation, then achievement motivation is a moderator variable.

Convergent and Discriminant Validity

Convergent validity exists when there is a high correlation between tests which measure the same trait. If two tests of verbal ability correlate .86, then we can say they exhibit convergent validity. Discriminant validity refers to the relative independence of tests measuring different traits.

As was illustrated above, the criterion for convergent validity is that the correlations among several tests measuring one trait must be significantly greater than zero. For discriminant validity, three criteria must be met:

1. The correlations of one trait (e.g., verbal ability) over several methods of evaluating that trait (e.g., test, teacher rating, peer rating) must be significantly greater than the correlations not having trait or method in common.

2. The correlations of one trait over several methods of evaluating that trait should be significantly higher than the correlations of different traits measured by the same method.

3. There should be a stable pattern of trait interrelationships regardless of the method used.

Many tests and measures which are now in use do not meet these criteria for convergent and discriminant validity. On the other hand, you as an evaluator should not undertake to solve this dilemma without the help of a statistician or psychologist. This is because the aforementioned criteria are relative rather than absolute, and because the analytic techniques are difficult to perform.

CHAPTER V

PRACTICAL APPLICATIONS

In Chapter V, we descriptively present examples of problems which have been investigated through application of some of the methods and procedures described earlier in this manual. The list of problems is selective rather than exhaustive. We have purposely picked, as illustrations, evaluation and measurement problems that are common. In most cases, we have generalized the description in order to emphasize research strategy rather than research specifics.

Cross-Cultural Training--Problem[1] *

The problem in this study done by the Navy was to determine whether or not subjects trained in a two-week Vietnamese language course could function as well as subjects in a six week Vietnamese language course.

The broad aim of most cross-cultural programs is to allow the serviceman to function behaviorally and effectively in a foreign environment. The present study attempted to evaluate different training methods in relationship to that aim.

Solution

A group of subjects was subdivided randomly into the two week and six week training conditions. This randomization procedure controlled for intelligence, motivation, and background factors across the two experimental groups. The two week language course was, simply, a shortened and condensed version of the six week language course. The data were analyzed via questionnaire and simple averaging. The results demonstrated that: (1) graduates of either course met most objectives in that they were able to acquire some vocabulary and conversational skills; (2) higher aptitude students performed extremely well in the six week course; (3) many graduates thought the six week course was inefficient; and (4) low aptitude students were only marginally adequate on graduation from the shorter course.

*All references in this chapter appear at the end of the chapter.

Emergency Training--Problem [2]

The problem investigated in this study was whether the use of "adjunct auto-instruction," a modification of programmed learning which keeps the learner active and gives him feedback, faci'. ates learning. The training was centered around pre-emergency preparation of the public for a disaster or critical situation.

Solution

The subjects in this study were four groups of matched semi-skilled, adult, employed women receiving attack survival material. The four experimental conditions were: (1) received material by telephone, (2) read material in print, (3) read material in print and received adjunct auto-instruction, (4) received material by telephone and received adjunct auto-instruction. The non-adjunct groups were presented the material twice to equate for exposure time. A final examination was administered at the end of training. An analysis of variance demonstrated that both adjunct trained groups were significantly superior in final learning level to the non-adjunct trained groups.

Questionnaire Measurement and Attitudes--Problem [3]

The problem in this study was to compare the confidence and attitudes of trainees taking "Quick Kill Basic Rifle Marksmanship" (QKBRM) training with the confidence and attitudes of trainees taking traditional Basic Rifle Marksmanship (BRM) training. QKBRM involved training the student to engage a target without aligning the sights of the weapon.

Many times investigators tend to measure achievement or performance in a vacuum. The attitudes and opinions of trainees should also be evaluated when comparing one training method with another. If you find that two training methods produce equivalent performance, then you would ordinarily select the method yielding superior attitudes. This can be complicated, though, by costs. Suppose that the training method yielding superior attitudes is more costly to implement. Your choice, then, rests with two factors: (1) how much extra cost you can absorb, and (2) the size of the difference in attitudes between both groups. If the cost differences are not large, one would probably use the training method producing better attitudes. If, on the other hand, the cost differences are large (relative to what you can afford), then the least costly method might be selected.

Solution

Two experimental groups received QKBRM in their training, and one matched control group received traditional BRM training. One of the experimental groups received pre and posttraining questionnaires, and the other experimental group received only a posttraining questionnaire. Control and experimental groups were compared on: (1) gains in confidence, (2) attitudes toward BRM, and (3) drill sergeant attitudes toward QKBRM. Significance tests indicated an increase in confidence in both groups with QKBRM trainees gaining more confidence than those trained through the BRM method. Instructor attitude to the QKBRM was less favorable than that of the trainees.

Comparative Evaluation--Problem [4]

Typically, comparative evaluation involves the comparison of an established training program with a new training program or innovative program. In this particular case, the investigators were interested in whether or not elimination of electronics theory training hampers the on-the-job performance of electronics technicians.

Solution

For this experiment, two groups of subjects were used. One group was trained in the conventional way, including instruction in electronics theory. The second group was trained in the usual manner, but without electronics theory. Each group consisted of 25 trainees. Control for chance differences in ability was exercised by matching the subjects assigned to each group on Army General Classification Test scores and on background characteristics. For control purposes, all aspects of both programs were the same except for the amount of electronics theory included in each. Any change in the dependent variable, then, was a result of the experimental manipulation and not the result of some uncontrolled factor.

Performance tests were employed to test posttraining performance. These are believed more relevant than written tests in this situation. These performance tests were chosen to reflect the job that the men were being trained for. The performance test scores of the two groups were compared using a special kind of analysis of variance. The results demonstrated that the non-theory trained group performed as well as the theory trained group. Accordingly, it was concluded that theory training is not necessary for electronics technicians in the situation involved.

Rater Bias--Problem[5]

When ratings are used in evaluative situations, it is very important to control for the rater's personal bias. These studies were undertaken because the checkpilot ratings were suspected to reflect their own standards rather than the student's flying skill.

Solution I

In a first study, the training program was analyzed into maneuver components. Proficiency scales and instrument observation were substituted for the checkpilot's own method. The Pilot Performance Description Record (PPDR) was constructed to reflect the most critical aspects of each maneuver. The PPDR was administered to 50 advanced and 50 intermediate level students. The results demonstrated:

1. improved reliability of flight proficiency evaluation

2. that the PPDR records specific student deficiencies

3. checkpilots who were trained in PPDR use were more consistent in their evaluation than checkpilots who were only oriented in its use.

Solution II

An objective and detailed scoring record, similar to the PPDR in "Solution I," was developed. Students were scored on checkrides during and after training. Class percentage errors were then calculated. This procedure allowed for class comparisons, grade comparisons, and instructor comparisons. If particular errors are identified among the students of one instructor, the instructor is given additional instructor training. Finally, if one checkpilot is more strict than the others, he is given counsel to make his observations more consistent with those of other checkpilots.

Training Low Aptitude Men--Problem[6]

This study was performed to determine the appropriateness of the selection standards for commisaryman training. The results of this study very clearly demonstrate the need for appropriate and realistic cut scores in any school selection situation.

Solution

Thirty-five low aptitude men underwent commisaryman training. As a control measure, another sample of non-low aptitude men also underwent the same training. An analysis of appropriate criterion data indicated that: (1) 31 of 35 low aptitude men successfully completed training, although their grades were significantly lower than the grades of non-low aptitude men, (2) low aptitude men needed to devote more outside time to study, (3) instructor interviews indicated that the low aptitude men required more time from instructors to meet course success criteria than non-low aptitude men, (4) analysis of variance showed that the differences between low aptitude men and non-low aptitude men were most evident on paper and pencil tests and least evident on actual job performance tests, (5) a correlational analysis demonstrated that Armed Forces Qualification Test (AFQT) scores failed to predict school performance, and (5) reading test scores were significantly correlated with some aspects of performance.

Individualized Training--Problem [7]

The problem in this classic study was to determine the impact of individualized training on the performance differences between low aptitude and non-low aptitude men in Basic Combat Training (BCT). This study also serves to illustrate the importance of the interaction concept.

Solution

High, middle, and low aptitude groups were selected, and individualized training was instituted using videotape, one to one student-teacher ratios, performance feedback, and small stepwise instructional increments. No control groups were used in this study. Learning time averages were recorded, and it was found that in some tasks low aptitude men reached standard, but took 2 to 4 times longer, and in other cases, they failed to master the material at all. Also, analyses of variance demonstrated that aptitude level interacted with method of instruction. The high aptitude group was found to learn equally well with lecture or with individualized training, while the low aptitude group learned well with individualized training, but not with the lecture method.

Intra-examiner Reliability--Problem [8]

In all evaluation situations involving performance tests, one should ascertain the reliability of the observers of the people who act as test administraters. The ideal method for determining the consistency of an individual examiner is the situation in which the examinee's performance is held constant over two separate occasions and the examiners' perceptions allowed to vary. Since the stimulus configuration remains constant, any unreliability shown can then be attributed to variation within the examiner. However, unfortunately, no one can possibly perform the same job in exactly the same manner on two separate occasions. One method by which performance may be held constant is to take a motion picture of the examinee performing the job. The motion picture may then be shown on two separate occasions and the examiner asked to score the action twice. Thus, the stimulus situation is held constant over the two time intervals, and any variations shown may be attributed to variation within the examiner. Two assumptions of this method are that the movie situation presents the same stimulus configuration to the examiner as does the actual work sample performance test situation and that the examiner scores the movie in the same manner as he would score an actual work sample performance test.

Solution

A 16 mm. black-white movie was made of a mechanic taking a Drill Point Grinding Work Sample Performance Test. The film was unrehearsed and the only instructions given the subject, a randomly selected mechanic, were "to grind the drill as he would ordinarily do it." The mechanic was told movies would be taken while he was working. The motion picture cameras and lights were not hidden, but their presence and the knowledge that his behavior was being photographed did not seem to affect the behavior of the mechanic.

The motion picture was then first shown to five supervisory mechanics. These supervisors had previous experience in work sample performance test administration and were moderately well informed in the general principles of work sample performance test administration. The movie was reshown to the same supervisors one month after its first administration. Therefore, each supervisor acted as his own control. One month is usually accepted as a sufficient time interval for forgetting of original responses. Moreover. the supervisors did not know that they would be asked to make exactly the same observations on two separate occasions. Therefore, there was little reason for them to try to remember their original responses.

The supervisors were asked to fill in a Movie Evaluation Form during each showing of the motion picture. The Movie Evaluation Form contained items such as: "Did the examinee check the tool rest for proper distance from the periphery of the grinding wheel? ";"Did the examinee ever adjust the tool rest while the grinding wheel was in motion? "; "Did the examinee wear loose clothing or clothing that could snag in the grinding wheel? "; "Did the examinee check the shank of the drill for bends and burns? "; etc. Sufficient light was allowed in the "theater" so that the supervisors could fill in the forms as the appropriate action was performed. Thus, the motion picture situation was as close as possible to actually scoring a work sample performance test.

A subject was considered consistent on an item if he answered the item on the second showing of the motion picture in exactly the same manner that he did on the first showing. Thus:

$$\text{Intra-examiner consistency} = \frac{\text{Number of items answered in exactly the same manner on each showing of motion picture}}{\text{Total number of items on questionnaire}} \times 100$$

The grand mean for intra-examiner agreement was 82.8 per cent, with a range from 64.3 per cent to 100 per cent. This mean of 82.8 per cent agreement would usually be considered adequate if converted into a correlation coefficient and interpreted as correlation coefficients are usually interpreted. Of course, these intra-examiner reliability estimates are based on only one motion picture. The danger of generalization from one measure of the reliability of observations of performance in process to all observations of performance in process is self evident.

In view of the range shown, the desirability of determining the reliability of the observations of examiners prior to assigning them to test administrative duties is also indicated. If all examiners show low consistencies, then either the examiner training has been poor or the test itself is inadequate. Naturally, only those examiners with high consistencies are worthy of consideration as test administrators.

The problem of how high examiner consistency must be before it is high enough remains open.

Interobserver Reliability--Problem [9]

Researchers have suggested that in a group performance test situation, scoring the intangible products of performance resulted in a slightly lower interobserver reliability than scoring the tangible products of performance. The present study compared interobserver consistency when the tangible and intangible products of performance on individually administered performance tests are scored. Since measurements of the final product (tangible measurements) can be made at the examiner's leisure, and since gauges and other measuring instruments can be employed as aids in making these estimates, some performance test constructors maintain that measurements of the final product should yield greater interexaminer consistency than measurements of performance in process (intangible products). An example of a measurement of performance in process (intangible product) is scoring the examinee's technique in doing a job, or scoring his adherence to the prescribed safety precautions. An example of a tangible product is the adherence of the final product to prescribed dimensions or standards. Since measurements of performance in process frequently tell where and how examinees erred rather than merely what mistakes were made, measurements of intangible products should be included in any performance test, provided there is no simultaneous loss in inter- (and intra-) observer consistency.

The present research was directed toward investigating the conjecture that for individually administered performance tests, final product measurement yields no greater interobserver consistency than observations of performance in process.

Solution

Eight performance tests, scored by the checklist method, were constructed. The checklists generally included items in the following four areas: (a) observations of the procedure followed in doing the job, (b) observations of the examinee's adherence to safety precautions, (c) observations of the examinee's methods of using tools and equipment, and (d) measurements of the final product. Items in the first three of these areas were intangible measurements, while items in the fourth area were tangible. Five of the tests were directed toward measuring the ability of mechanics and three were directed toward photographers. The mechanic's test included a rigid-tubing assembly test, a drill-point grinding test, a metal-working test, an aluminum butt-welding test, and a fabric-repair test. The tests for photographers included a motion-picture processing test, a test on the use of the Speed-Graphic camera, and a continuous-strip-printing test.

The mechanic's tests were administered to 19 examinees. Two supervisors independently, but simultaneously, administered the tests to each examinee. Examiner "A" acted as one of the test administrators throughout, and scored all 19 examinees. Examiners "B" through "F" each, simultaneously with "A," administered the battery to from three to five of these examinees. Thus, each examinee was independently evaluated by two examiners, examiner "A" and one other, both of whom simultaneously but independently, scored his work. Precautions were taken to insure that the examiners did not communicate with each other during the scoring, and each scoring sheet was collected immediately after an examiner was finished with it. Thus, even if an examiner decided on the basis of <u>post hoc</u> cross communication that he had erred, he was unable to correct his mistake. Moreover, the presence of the researcher served to enforce "security."

For the aerial photographers, a similar paradigm was used. One examiner (W) was constant throughout and administered all three photographic tests to a total group of 15 examinees. Each of the three other examiners (X, Y, or Z) independently, but simultaneously with "W," evaluated five of the 15 examinees.

An item analysis was first performed comparing Chief "A's" scoring of each item in a test area with the scoring of the identical item of the co-administrator who simultaneously, but independently, scored the same examinee. Percentage of consistency between simultaneous examiners for each test area was then obtained. Thus,

$$\begin{array}{c}\text{Interobserver consistency}\\ \text{within a test area}\end{array} = \frac{\begin{array}{c}\text{Number of items in test area scored in}\\ \text{same manner by simultaneous observers}\end{array}}{\text{Total number of items in area}} \times 100$$

Similarly, for the aerial photographers, it was possible to compare the agreement of examiner "W's" scorings with those of examiners "X," "Y," and "Z."

The results demonstrated that with regard to the mechanics, the mean interexaminer consistency was greater for all other measurements than it was for measurements of the final product. Similarly, for the aerial photographers, no regular superiority was seen in interexaminer consistency for measurements of the final product.

The reason for the comparatively high interexaminer reliabilities for the measurements in the intangible areas may be an outgrowth of the objectivity introduced into the checklist items and of the grossness of the observations called for in measuring performance in process (intangible products). If we want to know if someone is dead or alive, we can use a stethescope, but if the person is moving around, most observers will agree that he is alive without the use of the measuring instrument. Similarly, the observations of performance in process may have been gross enough and well defined enough to preclude the need for the stethoscope.

Transfer of Training--Problem [10]

Recently, technical training has shifted its focus from a general training to a training which introduces trainees directly into specialized instruction. It is the goal of this specialized training to produce men who are immediately useful upon graduation in a short "pipeline" time.

The newer program may be contrasted with the technical training previously given. This previous training was broader in nature, emphasized more deeply the theoretical aspects of the technical skills involved in maintenance, and relied to a greater extent on in-service training for imparting the specific technical skills needed for specific job performance. The problem in the present study was to compare in a real situation the technical effectiveness of technicians given specific training (transfer through identical elements) with the efficiency of trainees given a more general background knowledge (transfer through generalization).

Solution

Graduates of each training program for three naval ratings were studied: (a) jet aviation machinist's mate, (b) air controlman, and (c) parachute rigger. Success as a parachute rigger or a machinist's mate depends upon mechanical ability and perceptual motor skill, while success as an air controlman depends to a considerable degree on verbal behavior and abstract reasoning.

For each rating, a complete library or listing of the tasks that the technician could be called upon to perform in the fleet was developed and cast in technical behavior checklist (TBCL) form. Each checklist contained three parts. Since it was felt that one of the most valid indicators of acceptable performance is the willingness of a man's supervisor to assign him without direct, technical supervision to various technical tasks, Part I determined the amount of time spent by the ratee on each of the tasks within the rating. In Part II, the amount of supervision that the journeyman required on each of the tasks he performed was requested. Part III acquired an estimation of the criticality, in terms of squadron mission, of each of the tasks listed. In order to study the developmental aspects of the ratees, separate evaluations were obtained in Parts I and II for each of two time periods: the first three months the man being rated was in the fleet (T-1) and the fourth to ninth months he was in the fleet (T-2).

The subjects were graduates of the two naval aviation training programs, the previous general "A" school program and the more recently established and specialized program. For the air controlmen, 39 graduates of the more general training program and 42 graduates of the specialized program were studied on their fleet jobs.

For the parachute riggers, 10 graduates of the previous program and 23 graduates of the specialized training program were included. In the aviation motor machinist's mate rating, three groups were involved: graduates of the previous general training program (N = 21); graduates who had received an intermediate type of training which involved specialized training but which did not include training on the specific equipment used in the fleet (N = 60); and graduates of the specialized program who had received specialized training and practice on the specific equipment found in the fleet (N = 36). The subjects were distributed over six naval air stations and 32 squadrons.

In order to derive the criterion instrument, the mean time in hours spent during T-1 and T-2 by each group in each rating on each of the listed tasks was computed. For each task listed in Part II of the TBCLs, a score ranging from one to five was assigned; a score of five indicated proficient task performance after an initial checkout; a score of one was assigned for tasks on which the striker had received six or more checkouts but on which he was still unable to perform without direct supervision. Scores of four, three, or two were respectively assigned as follows: proficient after 3 to 5 checkouts, proficient after 6 or more checkouts, 1 to 5 checkouts but not proficient. The mean and variance of the Part II scores for each task on each of the three separate TBCLs were computed by time period. Additionally, overall means and overall variances were computed for each task in each of the three separate TBCLs. Last, the criticality of each task was derived from the responses in Part III of the TBCLs. Essentially, in developing the final criterion TBCLs, those tasks which were considered "unimportant" and which are relatively useless in a measurement instrument because of lack of discrimination (variability) between individuals were eliminated.

Thus, three separate final criterion TBCLs were developed. one for each of the ratings considered. The final criterion TBCL for motor machinists' mates contained 38 tasks; the final criterion TBCLs for parachute riggers and air controlmen contained 43 and 34 tasks respectively.

Using the same scoring method discussed above, Part II of each TBCL was rescored with the total score being equal to the sum of the scored tasks divided by the number of tasks attempted. The total scores so derived were then subjected to an analysis of variance.

The analysis of variance results enable a general answer to the question of whether the specialized training had exerted any general effects on fleet technical efficiency. For the aviation machinists' mates and parachute riggers no statistically significant between-groups differences were evidenced. For the air controlmen, the more generally trained group was superior to a statistically significant extent. Between time period differences were noted and were expected since a trainee would be expected to perform at a superior level during his fourth to ninth months in the service as compared with his first three months.

The next step was to determine where the men in the air controlman's rating were doing well and where they needed improvement. The response score distributions for the five task content clusters included in the final criterion air controlman's TBCL were derived. The distributions indicated that for each of the five content clusters (using equipment, using publications, testing equipment, receiving and transmitting messages, and controlling traffic), the generally trained group had higher means over both time periods. For the first time period both groups were poorest at the tasks involved in receiving and transmitting messages, while in the second time period the generally trained group was, by and large, proficient at these tasks, and the specifically trained group required additional training. Moreover, even in T-2, the specifically trained group remained weakest in receiving and transmitting messages and in controlling traffic. These tasks are believed to involve mostly nonroutine thinking as opposed to the specific information transfer involved in using publications, using equipment, and testing equipment.

REFERENCES FOR CHAPTER V

1 Abrams, A., & Pickering, E. An evaluation of two short vietnamese language courses. San Diego, Calif.: Naval Personnel and Training Laboratory, 1970.

2 Siegel, A., & Fischl, M. Mass training in civil defense: II. A further study of telephone adjunct training. Wayne, Pa.: Applied Psychological Services, 1965.

3 Olmstead, J. The effects of "quick kill" upon training confidence and attitudes. HumRRO Technical Report, No. 68-15, 1968.

4 Valverde, H. Learner Centered Instruction (LCI): A systems approach to electronic maintenance training. AMRL-TR-67-208, AD-846 721. Wright-Patterson AFB, Ohio: Air Force Medical Research Laboratories, 1968.

5 Greer, G., Smith, W., & Hatfield, J. Improved flight proficiency evaluation in Army helicopter pilot training. HumRRO Technical Report, No. 69-1, 1969.

6 Duffy, J., & Jolley, O. Briefing on task LIFT. Collected papers prepared under work unit LIFT: Army Aviation Helicopter Training. HumRRO Professional Paper, No. 18-68, 1968, 3-10.

7 Hooprich, E. A second investigation of feasibility of Navy commissary man training for group IV personnel. San Diego, Calif.: Naval Personnel Research Activity, 1968.

8 McFann, H. Individualization of Army training. Innovations for training. HumRRO Professional Paper, No. 6-69, 1969, 1-9.

9 Siegel, A. Retest-reliability by a movie technique of test administrators' judgments of performance in process. Journal of Applied Psychology, 1954, 38(6), 390-392.

10 Siegel, A. Interobserver consistency for measurements of the intangible products of performance.' Journal of Applied Psychology, 1955, 39(4), 280-282.

11 Siegel, A., Richlin, M., & Federman, P. A comparative study of "transfer through generalization" and "transfer through identical elements" in technical training. Journal of Applied Psychology, 1960, 44(1), 27-30.

SELECTED READING LIST

CHAPTER I

Campbell, J. Personnel training and development. Minneapolis, Minn.: Minnesota University Department of Psychology, 1970.

Cronbach, L. Course improvement through evaluation. Teachers College Record, 1963, 64, 672-683.

Furno, O. Sample Survey designs in education-focus on administrative utilization. Review of Educational Research, 1966, 36, 552-565.

Gagné, R. (Ed.) Psychological principles in system development. New York: Holt, Rinehart & Winston, 1962.

Glaser, R., & Glanzer, N. Abstract of training and training research. Pittsburgh: American Instiiutes for Rese rch, 1958.

Hawkridge, D. Designs for evaluative studies. In Evaluative research: Strategies and methods. Pittsburgh: American Institutes for Research, 1970. Pp. 24-47.

Hunter, N., Lyons, J., MacCaslin, E., Smith, R., & Wagner, H. The process of developing and improving course content for military technical training. Alexandria, Va.: George Washington University, Human Resources Research Office, 1969.

Instructional System Development. Washington, D.C.: Department of the Air Force, Air Force Manual 50-2, 1970.

Osborn, W. An approach to the development of synthetic performance tests for use in training evaluation. Alexandria, Va.: HumRRO Professional Paper, No. 30-70, AD-719 265, 1970.

Tyler, R. (Ed.) Educational evaluation: New roles, new means. Chicago: University of Chicago Press, 1969.

Wittrock, M., & Wiley, D. (Eds.) The evaluation of instruction. New York: Holt, Rinehart & Winston, 1970.

Preceding page blank

CHAPTER II

Angell, D., Shearer, J., & Berliner, D. Study of performance evaluation techniques. Port Washington, N.Y.: U.S. Naval Training Device Center, 1964.

Bass, B. Thiagarajan, K., & Ryterband, E. On the assessment of training value of small group exercises for managers. Rochester, N.Y.: University of Rochester, 1968.

Campbell, J. Personnel training and development. Minneapolis, Minn.: Minnesota University, Department of Psychology, 1970.

Campbell, J. Personnel training and development. In P. Mussen & M. Rosenzweig (Eds.), Annual Review of Psychology. Palo Alto, Calif.: Annual Reviews Inc., 1971, 565-602.

Cronbach, L. Course improvement through evaluation. Teachers College Record, 1963, 64, 672-683.

Schultz, D., & Siegel, A. Post-training performance criterion development and application: A selective review of methods for measuring individual differences in on-the-job performance. Wayne, Pa.: Applied Psychological Services, 1961.

Smode, A., Hall, E., & Meyer, D. An assessment of research relevant to pilot training. AMRL-TR-66-99, AD-803 281. Wright-Patterson AFB, Ohio: Aerospace Medical Research Laboratories, 1966.

Suchman, E. Evaluative research: Principles and practice in public service and social action programs. New York: Russell Sage Foundation, 1967.

Tyler, R. (Ed.) Educational evaluation: New roles, new means. Chicago: University of Chicago Press, 1969.

Walker, R. An evaluation of training methods and their characteristics. Human Factors, 1965, 7(4), 347-354.

Wittrock, M., & Wiley, D. (Eds.) The evaluation of instruction. New York: Holt, Rinehart & Winston, 1970.

CHAPTER III

Carver, R. The curvilinear relationship between knowledge and test performance: Final examination as the best indicant of learning. In K. Wientge & P. DuBois (Eds.) *Criteria in learning research.* St. Louis, Mo.: Washington University, Technical Report No. 9, 1966.

Carver, R. A model for using the final examination as a measure of the amount learned in classroom learning. *Journal of Educational Measurement,* 1969, *6,* 59-68.

Carver, R. Special problems im measuring change with psychometric devices. In *Evaluative research: Strategies and methods.* Pittsburgh: American Institutes for Research, 1970. Pp. 48-66.

Cleary, T., Linn, R., & Rock, D. Reproduction of total test scores through the use of sequentially programmed tests. *Journal of Educational Measurement,* 1968, *5,* 183-187.

Cleary, T., Linn, R., & Rock, D. An exploratory study of programmed tests. *Educational and Psychological Measurement,* 1968, *28,* 345-360.

Comrey, A. A proposed method for absolute ratio scaling. *Psychometrika,* 1950, *15,* 317-325.

Coombs, C. Milholland, J. & Warner, F. The assessment of partial knowledge. *Educational and Psychological Measurement,* 1956, *16,* 13-37.

Denova, C. Is this any way to evaluate a training activity? You bet it is! *Personnel Journal,* 1968, *4(7),* 488-493.

DuBois, P. *Multivariate correlational analysis.* New York: Harper & Brothers, 1957.

Duffy, J., & Jolley, O. Briefing on task LIFT. Collected papers prepared under work unit LIFT: Army Aviation Helicopter Training. *HumRRO Professional Paper,* No. 13-68, 1968, 3-10.

Echternacht, G., Sellman, W., Bold., R., & Young, J. *An evaluation of the feasibility of confidence testing as a diagrostic aid in technical training.* AFHRL-TR-71-33, AD-734 032. Lowry AFB, Colo.: Technical Training Division, Air Force Human Resources Laboratory, July 1971.

Englemann, S. Relating operant techniques to programming and teaching. *Journal of School Psychology*, 1968, 6(2), 89-96.

Ferguson, R. *Computer-assisted criterion-referenced testing.* Pittsburgh: Learning Research and Development Center, University of Pittsburgh, 1969.

Ferguson, R. *Computer-assisted criterion-referenced measurement.* Pittsburgh: Learning Research and Development Center, University of Pittsburgh, 1970.

Fitzpatrick, R. The selection of measures for evaluating programs. In *Evaluative research: Strategies and methods.* Pittsburgh: American Institutes for Research, 1970. Pp. 67-81.

Fleishman, E. (Ed.) *Studies in personnel and industrial psychology.* Homewood, Ill.: The Dorsey Press, 1967.

Gagné, R. Curriculum research and the promotion of learning. In *Perspectives of curriculum evaluation.* Chicago: Rand McNally, 1967. Pp. 19-38.

Glaser, R., & Cox, R. *Criterion-referenced testing for the measurement of educational outcomes.* Pittsburgh: Learning Research and Development Center, University of Pittsburgh, 1968.

Glaser, R., & Nitko, A. *Measurement in learning and instruction.* Pittsburgh: Learning Research and Development Center, University of Pittsburgh, 1971.

Gronlund, N. *Constructing achievement tests.* Englewood Cliffs, N.J.: Prentice-Hall, 1968.

CHAPTER III (cont.)

Jacobs, S. Correlates of unwarranted confidence in responses to objective test items. *Journal of Educational Measurement*, 1971, *8(1)*, 15-20.

Johnson, K. *Identification of difficult units in a training program.* San Diego, Naval Training Research Laboratory, Technical Bulletin, STB 69-4, 1969.

Lawshe, C., & Balma, M. *Principles of personnel testing.* New York: McGraw-Hill, 1966.

Osgood, C., Suci, G., & Tannenbaum, P. *The measurement of meaning.* Urbana, Ill.: University of Illinois Press, 1957.

Popham, W., & Husek, J. Implication of criterion-referenced measurement. *Journal of Educational Measurement*, 1969, *6*, 1-9.

Principles and techniques of instruction. Washington, D. C.: Department of the Air Force, AF Manual 50-9, 3 April 1967.

Shuford, B. *Confidence testing: A new tool for measurement.* Lexington, Mass.: Shuford-Massengill Corp., 1967.

Siegel, A., & Fischl, M. *Absolute-scales of electronic job performance: Empirical validity of an absolute scaling technique.* Wayne, Pa.: Applied Psychological Services, 1965.

Siegel, A., Schultz, D., & Lanterman, R. *The development of absolute scales of electronic job performance.* Wayne, Pa.: Applied Psychological Services, 1964.

Siegel, A., & Schultz, D. Evaluating the effects of training. *Journal of American Society of Training Directors*, 1961.

Torgerson, W. *Theory and methods of scaling.* New York: Wiley. 1958.

Tiffin, J., & McCormick, E. *Industrial Psychology.* Englewood Cliffs, N. J.: Prentice-Hall, 1965.

Wood, D. *Test construction.* Columbus, Ohio: Merrill, 1960.

CHAPTER IV

Bruning, J., & Kintz, B. *Computational handbook of statistics.* Glenview, Ill.: Scott, Foresman, 1968.

Campbell, D., & Fiske, D. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin,* 1959, *56,* 81-105.

Campbell, D., & Stanley, J. *Experimental and quasi-experimental designs for research.* New York: Rand McNally, 1963.

Guilford, J. *Fundamental statistics in psychology and education.* New York: McGraw-Hill, 1965.

Harman, H. *Modern factor analysis.* Chicago: University of Chicago Press, 1967.

Hays, W. *Statistics for psychologists.* New York: Holt, Rinehart & Winston, 1963.

Siegel, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

Winer, B. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

CHAPTER V

Abrams, A., & Pickering, E. *An evaluation of two short vietnamese language courses.* San Diego: Naval Personnel and Training Laboratory, 1970.

Duffy, J., & Jolley, O. Briefing on task LIFT. Collected papers prepared under work unit LIFT: Army Aviation Helicopter Training, *HumRRO Professional Paper,* No. 18-68, 1968, 3-10.

Greer, G., Smith, W., & Hatfield, J. Improving flight proficiency evaluation in army helicopter pilot training. *HumRRO Technical Report,* No. 69-1, 1969.

Hooprich, E. *A second investigation of the feasibility of navy commissaryman training for group IV personnel.* San Diego: Naval Personnel Research Activity, 1968.

McFann, H. Individualization of army training. Innovations for training. *HumRRO Professional Paper,* No. 6-69, AD-685 498, 1969, 1-9.

Olmstead, J. The effects of "quick kill" upon training confidence and attitudes. *HumRRO Technical Report,* No. 68-15, AD-682 350, 1968.

Siegel, A. Interobserver consistency for measurements of the intangible products of performance. *Journal of Applied Psychology,* 1955, *39(4),* 280-282.

Siegel, A. Retest-reliability by a movie technique of test administrator's judgements of performance in process. *Journal of Applied Psychology,* 1954, *38(6),* 390-392.

Siegel, A., & Fischl, M. *Mass training techniques in civil defense: II. A further study of telephone adjunct training.* Wayne, Pa.: Applied Psychological Services, 1965.

Siegel, A., Richlin, M., & Federman, P. A comparative study of "transfer through generalization" and "transfer through identical elements" in technical training. *Journal of Applied Psychology,* 1960, *44(1),* 27-30.

INDEX